Received: 8 June 2020 | Accepted: 31 August 2020 | Published Online: 8 October 2020
DOI: 10.36850/mr1

Burst Beliefs – Methodological Problems in the Balloon Analogue Risk Task and Implications for Its Use

JOTE
Journal of Trial and Error

Kristel De Groot[1]

[1]Erasmus University Rotterdam Institute for Behaviour and Biology (EURIBEB) | Institute of Psychology, Erasmus School of Social and Behavioural Sciences | Department of Applied Economics, Erasmus School of Economics; Erasmus University Rotterdam, the Netherlands

Correspondence
Kristel De Groot, Department of Psychology, Education & Child Studies/ Research Methods and Techniques, Erasmus University, Rotterdam, the Netherlands
Email: k.degroot@ese.eur.nl

Funding:
The author is supported by the Research Talent programme with project number 406.17.505, which is financed by the Dutch Research Council (NWO).

Acknowledgments: The author would like to acknowledge Jan van Ours, Marco Lauriola, and Sander Wieman for their constructive feedback on previous versions of this manuscript, and for sharing their views on the strengths and weaknesses of the BART. The author also expresses gratitude to Sander Wieman for his thorough language-editing and for proofreading the final manuscript.

Abstract

Studies in the field of psychology often employ (computerized) behavioral tasks, aimed at mimicking real-world situations that elicit certain actions in participants. Such tasks are for example used to study risk propensity, a trait-like tendency towards taking or avoiding risk. One of the most popular tasks for gauging risk propensity is the Balloon Analogue Risk Task (BART; Lejuez et al., 2002), which has been shown to relate well to self-reported risk-taking and to real-world risk behaviors. However, despite its popularity and qualities, the BART has several methodological shortcomings, most of which have been reported before, but none of which are widely known. In the present paper, four such problems are explained and elaborated on: a lack of clarity as to whether decisions are characterized by uncertainty or risk; censoring of observations; confounding of risk and expected value; and poor decomposability into adaptive and maladaptive risk behavior.

Furthermore, for every problem, a range of possible solutions is discussed, which overall can be divided into three categories: using a different, more informative outcome index than the standard average pump score; modifying one or more task elements; or using a different task, either an alternative risk-taking task (sequential or otherwise), or a custom-made instrument. It is important to make use of these solutions, as applying the BART without accounting for its shortcomings may lead to interpretational problems, including false-positive and false-negative results. Depending on the research aims of a given study, certain shortcomings are more pressing than others, indicating the (type of) solutions most needed. By combining solutions and openly discussing shortcomings, researchers may be able to modify the BART in such a way that it can operationalize risk propensity without substantial methodological problems.

KEYWORDS
Balloon Analogue Risk Task, risk-taking, uncertainty, expected value, confounding, censoring.

Purpose

The Balloon Analogue Risk Task (BART) is one of the most widely used behavioral tasks in psychology and has an especially strong presence in the fields of decision research, addiction research, and neuropsychology. But despite its popularity, researchers using the BART seem largely unaware of the task's methodological shortcomings, which sometimes leads to conclusions that are not supported by the data. This is likely a result of these shortcomings not being widely reported, as 'failure' is not considered a popular publishing theme. Therefore, the present paper aims to gather and review these shortcomings, as well as potential solutions.

Take-home Message

The Balloon Analogue Risk Task (BART) suffers from various methodological shortcomings. The present paper analyses these shortcomings and offers suggestions to mitigate their effects. Finally, it calls upon researchers to critically evaluate how these shortcomings impact their studies before deciding whether and how to use BART.

Introduction

To a large extent, psychological science rests on the promises of operationalization: defining fuzzy concepts as measurable variables, or in other words, changing conceptual variables into operational ones (Shuttleworth, 2008). This

process is imperative because most concepts researchers hypothesize about are not straightforwardly quantifiable. By defining how a concept is measured, operationalization allows hypotheses to take a falsifiable format and enables us to replicate findings. In a way, operationalizations are arbitrary, as concepts can be defined and thus measured in numerous ways – none of which are surely 'right'. Nonetheless, some measures may be more suitable than others.

A notable example of a concept that can be operationalized in various ways is risk-taking (Lauriola & Weller, 2018), which has an important place in clinical, cognitive, and developmental psychology, as well as in the fields of criminology, economics, and management. One way risk-taking is operationalized in these fields is through self-report measures, such as the Domain-Specific Risk-Taking (DOSPERT) scale (Blais & Weber, 2006) and the Financial Risk Tolerance assessment (Grable & Lytton, 1999). Another way is through computerized behavioral tasks, like the Iowa Gambling Task (Bechara et al., 1994), the Cambridge Gambling Task (Rogers et al., 1999), the Game of Dice Task (Brand et al., 2005), the Balloon Analogue Risk Task (Lejuez et al., 2002), and the more recent but already widely used Columbia Card Task (Figner et al., 2009). Importantly, the quality of a study largely depends on the degree to which its operational measures reflect the underlying concept; in this case, one's disposition towards risk-taking. If a task is a poor proxy for a concept or is subject to methodological or interpretational problems, any data resulting from it are of limited value to our understanding of the concept. In this regard, several studies have challenged the operationalization ability of the most-cited risk task, the Iowa Gambling Task (see e.g. Brand et al., 2006; Buelow & Suhr, 2009; Figner et al., 2009). The Balloon Analogue Risk Task, which is the second-most cited, may yet suffer from even more severe issues, hindering its ability to operationalize risk-taking. While some individual issues have been reported in previous publications, no literature so far has discussed these collectively. The present commentary aspires to fill this gap.

The Balloon Analogue Risk Task

In the Balloon Analogue Risk Task, or BART for short, participants are presented with a computer screen showing a small balloon and a pump. They are told that every time they click the pump, the balloon expands, and a fixed amount of money (5 cents) is added to a temporary bank. Every pump also increases the chance of the balloon exploding (marked by a 'pop' sound from the computer), resulting in losing all money in the temporary bank for that particular balloon (trial). The point at which a balloon explodes varies across trials, ranging from the first pump to the point where the balloon fills the entire screen. Participants can decide to stop pumping the balloon at any point during a trial by clicking the 'collect' button (left in Figure 1), which transfers the money accumulated in their temporary bank to their permanent one, while a slot machine sound is played. Once a balloon explodes or once participants cash a balloon's proceeds, the trial ends, and a new, uninflated, balloon appears.

In the original study by Lejuez et al. (2002), participants were informed that they would complete 90 balloons: 30 orange, 30 yellow, and 30 blue ones. Unbeknownst to participants, differently colored balloons had a different chance of exploding. The probability distribution governing their explosion points consisted of an array of n numbers from which on every pump a random number was drawn without replacement. If a 1 was drawn, the balloon exploded.

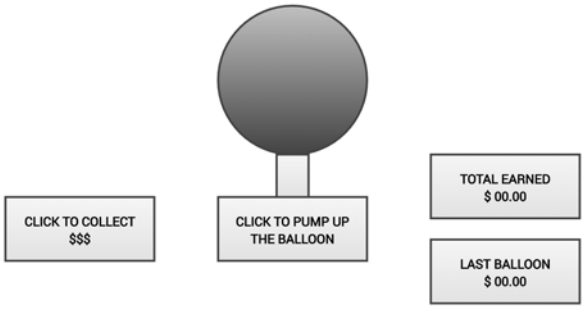

FIGURE 1 Set-up of the original Balloon Analogue Risk Task as described by Lejuez et al. (2002). An interactive illustration of the task is provided with the HTML version of this article.

Thus, the probability p of the balloon exploding on the first pump was $1/n$, and the probability of it exploding on pump i (given no prior explosion) was $p_i = \frac{1}{n-i+1}$. For orange balloons, the array ranged from 1 to 8 (hence $p_1 = \frac{1}{8-1+1} = 1/8$), for yellow balloons from 1 to 32 ($p_1 = \frac{1}{32-1+1} = 1/32$), and for blue ones from 1 to 128 ($p_1 = \frac{1}{128-1+1} = 1/128$). Their average explosion points were respectively 4, 16, and 64, with the same (randomly generated) sets of explosion points being used across all participants to limit extraneous variability. Neither the ranges nor the average explosion points were communicated to participants.

The BART's design is intended to reflect naturalistic decision-making, in which taking more risk generally increases the odds of encountering a loss. This sort of decision-making tends to be emotionally engaging, instigating a sense of increasing tension as the balloon increases in size (Schonberg et al., 2011). In support of the BART's validity, Lejuez et al. (2002) showed that the average number of times participants pumped the blue balloon significantly correlated with scores on risk-related constructs (sensation seeking, impulsivity) and with real-world risk behaviors, such as polydrug use, gambling, unsafe sex, and stealing. The orange and yellow pumps were originally not examined with respect to risk-related constructs, as their narrow ranges of outcome values (1-8 and 1-32) are less suited for capturing individual differences. Instead, their average pump numbers were analyzed together with those of the blue balloons to show that the number of times participants choose to pump is sensitive to the probability of exploding. Overall, the data showed the BART to have "particular promise as a behavioral index of risk-taking" (Lejuez et al., 2002, p. 82). As would be expected based on this conclusion, the BART (particularly its blue balloon) became a popular instrument for gauging individuals' propensity for risk-taking, with inconsistent findings being attributed to factors like sampling variability and inadequate statistical power (Lauriola et al., 2014), rather than problems inherent to the BART. However, several authors have argued that such problems exist (De Groot & Thurik, 2018; Gu et al., 2018; Schmidt et al., 2019), and that they limit the BART's ability to measure one's propensity for taking risk. The key problems that characterize the BART are 1) a lack of clarity as to whether decisions are characterized by uncertainty or risk, 2) censoring of observations, 3) confounding of risk and expected value, and 4) poor decomposability into adaptive and maladaptive risk behavior.

Risk or Uncertainty?

In economic theories of decision-making, a key distinction is that between uncertainty and risk, which is often accredited to Knight (1921), and was introduced to psychological thinking in a seminal paper by Edwards (1954) that lies at the origin of behavioral decision theory. When deciding under the condition of risk, the probabilities associated with the possible outcomes are *known*. When deciding under uncertainty (which some authors call ambiguity), this probability distribution is *unknown*.

For Knight (1921), this distinction was not only of theoretical but of practical importance as well. According to him, uncertainty – not risk – was the main driver of entrepreneurial success, as only people who recognize hidden opportunities can seize them and profit from them. Since then, the empirical relevance of the uncertainty-risk distinction has been confirmed in various fields of research. In economics, Ellsberg (1961) showed that individuals prefer risk over uncertainty, even if the known probabilities are unfavorable and the uncertain option could be a guaranteed win. In psychology, studies showed that uncertain and risky decisions involve different mental processes, as risk allows for statistical thinking (to optimize) but uncertainty involves heuristics (to satisfice) (Volz & Gigerenzer, 2012). In line with this, decision-making under risk is thought to depend more on executive function (such as categorization and cognitive flexibility) for which the dorsolateral prefrontal cortex is important, whereas decision-making under uncertainty hinges on emotional processes (such as somatic feedback), which are more associated with the ventromedial prefrontal cortex and the amygdala (Brand et al., 2006). This may explain why patients with executive deficits, such as those with Parkinson's disease, have difficulty deciding under risk but have no trouble deciding under uncertainty (Euteneuer et al., 2009), whereas persons with obsessive-compulsive disorder, for example, show the opposite pattern (Starcke et al., 2009, 2010).

Given that uncertainty and risk differ both theoretically and empirically, it is imperative for researchers to know the conditions under which participants decide. Unfortunately, despite the word 'risk' in its name, these conditions are not straightforward in the BART. Since participants are never given "detailed information about the probability of an explosion" (Lejuez et al., 2002, p. 77), we can assume that at least during early trials, they decide under uncertainty (Bishara et al., 2009; De Groot & Thurik, 2018; Schonberg et al., 2011). As they move further along in the task and 'sample the distribution' by pumping balloons and observing their outcomes, they get a better sense of the probabilities, which gradually moves their decisions in the direction of risk. Although not studied in the BART itself, such a shift has been shown for the Iowa Gambling Task, where performance in early trials does not correlate with that in later trials nor with executive function, indicating that people first decide under uncertainty and later under risk (Brand et al., 2006; Brand et al., 2007). While this effect may not be as strong in the BART, studies do show better performance in later compared to early trials, suggesting that participants indeed get a better grasp of the probability distribution over time (De Groot & van Strien, 2019; Lejuez et al., 2002).[1]

The BART's transition from uncertainty towards risk is problematic for several reasons. First, it is unclear *when* exactly this shift transpires, making it difficult to determine whether a decision in a given trial is made under uncertainty, risk, or something in between. Second, the point where decisions shift from uncertainty to risk is likely to differ between individuals, and is dependent on task characteristics (Brand et al., 2006; Brand et al., 2007). Third, the shift implies that the BART imposes learning demands, which could inadvertently impact participants' outcomes on the task, with those capable of updating their knowledge of the probabilities performing better than those who have difficulty doing so. Fourth, once participants manage to derive the task's probabilities, subsequent decisions are not characterized by what is usually considered risk. Contrary to decisions in which probabilities are explicitly described ('a priori' probabilities), probabilities in the BART are derived from experience. Since such probabilities depend on factors like sampling variability and one's memory of previous events, decision-makers treat experience-based probability differently, which is called the description-experience gap (Hau et al., 2008; Rakow & Newell, 2010). Most notably, when deciding based on experience, people do not act in accordance with prospect theory, but instead, underweight rare events and overweight common encounters. As people have more and more encounters (e.g. trials), their experiences will approach the precision of a priori probabilities, though in practice this is difficult to attain (Knight, 1921).

To address the inability of the BART to differentiate between complete uncertainty, experience-based risk, and description-based risk, several approaches may be used. One option is to apply a model to the BART's data that allows for participants learning through experience. An early example is a model by Wallsten et al. (2005) in which decision-makers update their probabilities from trial to trial, and continually re-evaluate their options. Alternatively, one could use a different task, in which decisions are either all characterized by uncertainty or risk, or which includes a well-understood shift between the two. Tasks that involve only uncertain decision-making are rather difficult to design, as they require participants to be ignorant of probability-related information and *remain* ignorant of that as well – automatically disqualifying tasks that have a learning curve. Tasks involving only decisions made under (a priori) risk are much more common and include the Cambridge Gambling Task, the Game of Dice Task, and the Columbia Card Task, the latter of which resembles the BART's dynamic, affective nature (Schonberg et al., 2011). Finally, a known shift from uncertainty to (experience-based) risk can be found in the Iowa Gambling Task. This task's shift, while not *fully* understood, has been studied more thoroughly than that in the BART.

Censored Observations

Statistical censoring refers to a condition in which the value of an observation is unknown because it is beyond a certain limit. This limit can exist by design, which is common in survival analysis. If a study on a surgical intervention follows patients for up to 10 years, the longevity scores of those who live past this term are censored, as their longevity is *at least* 10 (Young & McCoy, 2019). Censoring can also result from limits on what an instrument can reliably measure. For example, the full IQ score of the Wechsler Adult Intelligence Scale ranges from 40 to 160 (Sattler & Ryan, 2009), meaning that IQ scores of people performing either extremely poorly or extremely well are cut off at these boundaries and are thus censored.

1. The relevant data collected by De Groot and van Strien, 2019 on per-block averages is not reported in the published report but will be shared upon request.

In the BART, censoring (by design) occurs if a participant is stopped from taking more risk in a given trial, because the balloon they are pumping explodes, forcing the trial to end. Since such a trial ends prematurely, the number of times the participant pumped the balloon does not necessarily reflect the risk they were willing to take, meaning their risk propensity is censored. This is problematic for various reasons. First, including these censored trials biases the average number of pumps downwards (especially for high-risk takers), underestimating participants' willingness to take risks (Dijkstra et al., 2020; Pleskac et al., 2008). Likewise, the between-subjects variability across these averages is reduced (Lejuez et al., 2002). Overall, the (unadjusted) average number of pumps is an ill-suited operationalization of risk propensity.

As censoring affects all sequential risk-taking tasks like the BART (involving multiple decisions per trial) and various other research paradigms, like survival analysis, several solutions have been proposed. In the paper introducing the BART, Lejuez et al. (2002) suggest computing an adjusted pump average using only trials in which participants stopped voluntarily, that is, in which the balloon did not burst. However, by omitting explosion trials, censored observations are essentially treated as randomly missing, which is inaccurate (Pleskac et al., 2008). The more risk someone takes, the more likely it is that the balloon bursts, and that the trial forcibly ends. The termination of trials is therefore not independent from participants' behavior. As a result, Lejuez et al.'s adjusted score tends to discard trials in which participants take a lot of risk. This causes the average number of pumps to be biased downwards, similar to the unadjusted score, but to a lesser extent.

To circumvent the problem of censoring, Pleskac et al. (2008) developed an automatic response version of the BART.[†] Contrary to the standard BART, in which participants inflate a balloon one pump at a time, the automatic BART lets them indicate their intended number of pumps beforehand. The balloon then inflates to the corresponding size, or until it bursts. This procedure allows for an unbiased statistic of risk propensity, as the intended number of pumps is now observable in all trials (Pleskac et al., 2008). However, it increases the time between decision and outcome, which may make decisions less emotional (impulsive) and more cognitive (planned) (Pleskac et al., 2008), and may reduce the salience of the outcomes. These effects, in turn, can affect participants' risk-taking (Young & McCoy, 2019). In contrast, however, a study using the Bomb Risk Elicitation Task (BRET; Crosetto & Filippin, 2013), another risk task that uses delayed explosions to circumvent censoring, found that introducing such delays did not impact risk-taking.

Another solution to censoring is using a rigged task (Slovic, 1966). Participants are then told that failure can occur at any moment (in the BART, at any pump), but actually, it is set to occur at the last possible choice. Hence, participants can always stop voluntarily, and no scores are censored. To uphold credibility, 'mock' trials are added, in which failure is set to occur early on. Deciding on the number and timing of mock trials, however, is a challenge. Since behavior in a trial is affected by previous outcomes, experiencing (too) few failures could increase risk-taking (De Groot & van Strien, 2019; Dijkstra et al., 2020). Therefore, rigged tasks should be designed such that they produce failure rates similar to non-rigged tasks and should take into account that failure rates differ between participants too. However, research on the Columbia Card Task, another sequential risk-taking task, shows that this is often not the case (De Groot & van Strien, 2019).

A final remedy, which addresses the bias but leaves the BART unchanged, is to apply a statistical model to the resulting data that explicitly incorporates censored behavior. Such models consider *all* observed data, using the censored trials as lower bounds in determining a participant's actual risk propensity. Some of them employ Bayesian (generalized) linear mixed-effects regression (Weller et al., 2019; Young & McCoy, 2019); others use maximum likelihood estimation, adding a cumulative distribution function to the likelihood function to account for censoring (Dijkstra et al., 2020; Tobin, 1958). Such models perform significantly better (i.e., have less biased predictions) than those that do not account for censoring. However, as is the case for all statistical models, their soundness hinges on the validity of their underlying assumptions (Schafer & Graham, 2002), such as that of normality, whose violation not all models are robust against (Powell, 1984).

Confounding and Decomposability

The BART was designed to resemble real-world risk situations, where taking modest risk is generally advantageous, but taking excessive risk is increasingly unfavorable (Lejuez et al., 2002; Wallsten et al., 2005). Within a trial, every successful pump earns participants 5 cents, which are added to their temporary bank. As the amount accumulated in the bank grows, the relative gain of taking additional risk decreases, while the potential loss in case of an explosion increases. Additionally, the probability of the balloon exploding increases with every pump: from 1/128 on the first to 1/127 on the second, and so on.

This combination of characteristics makes that the task's structure entails a serious problem. Since both the balloon value (the amount collected in the temporary bank) and the explosion probability increase with every pump, the expected value of inflating the balloon – the product of the success chance and the reward, minus the product of the explosion chance and the balloon value – changes across a trial (Schmidt et al., 2019). This change is illustrated in Table 1. Early in a trial, the expected value of the pump is positive, so taking additional risk is advantageous. This prospect changes halfway when the expected value turns negative, making additional pumps unfavorable (Lejuez et al., 2002). Due to the expected value changing with each decision, it is *confounded* with risk (defined as the variability of the possible outcomes), which varies across decisions by design. Although such confounding can happen in real-life decision-making, it is not desirable in a controlled scientific environment: it makes it difficult to measure participants' risk propensity, as both risk and expected value may influence their decisions. The extent to which individuals are, for example, risk-seeking, can therefore not be determined, because this would require showing a preference for higher variance payoffs, holding expected value constant (Schonberg et al., 2011).

This confounding demonstrates that the BART's main observable outcome – the number of pumps participants press – cannot be interpreted as a straightforward indicator of risk propensity. Like many behavioral tasks, the BART supposedly gauges a single cognitive construct, but it manipulates various other, potentially confounding constructs as well (Schonberg et al., 2011). Expected value is an example of such a construct. As a result, the single score provided by the BART cannot easily be decomposed to identify the cognitive or neural

†. An interactive illustration of this task is provided with the HTML version of this article.

mechanisms involved in the pump decisions. Studying the risk-taking process in isolation using the BART is therefore not possible.

One approach for resolving the confounding and decomposability issues in the BART is to apply a computational model to its data that quantifies the cognitive mechanisms underlying the observed behavior (Bishara et al., 2009). Such models were first proposed by Wallsten et al. (2005), inspired by an expectancy-valence model for decomposing behavior in the Iowa Gambling Task (Busemeyer & Stout, 2002). Wallsten et al. explain decision variability using one parameter for risk-taking, one for response consistency, and two for learning. By applying these models, we can study risk-taking – and other aspects that determine BART behavior – in isolation, by translating "what is observed but relatively uninformative to what is unobserved and relatively informative" (van Ravenzwaaij et al., 2011, p. 95). However, data from the BART may not be rich enough to warrant the use of complicated decomposition models. For instance, a study on Wallsten et al.'s best performing model demonstrated that its learning parameters could not reliably be recovered (van Ravenzwaaij et al., 2011). To allow for more extensive decomposition, one may need to resort to a different task, like the Iowa Gambling Task. Alternatively, one could use a task that by design avoids confounding, such as the Columbia Card Task. Although dynamic and affective like the BART, this task orthogonally varies risk-related constructs, so that they can be decomposed into their underlying mechanisms – like sensitivity to gains, losses, and probabilities – without the use of a computational model (Dijkstra et al., 2020; Figner et al., 2009; Schonberg et al., 2011). Finally, researchers can choose to design a custom task to ensure that the constructs relevant to their hypotheses are not confounded. For example, a risk task presented in Schmidt et al. (2013) varies the level of risk but holds expected value constant. Solutions such as these should be considered carefully so that constructs crucial to a study's hypotheses can be isolated effectively.

The Normative Solution

The BART is designed in such a way that the balloons' average explosion point lies at 64, halfway the maximum number of pumps. This is achieved by randomly generating collections of explosion points until one produces an average of 64 over all trials, as well as within each set of 10 trials (Lejuez et al., 2002). Participants can then maximize their earnings by attempting to pump every balloon 64 times, which results in an explosion in about half of the trials, and an optimal overall expected value. Going back to Table 1, we can see exactly why this is the optimal, or *normative*, solution in the BART. Up to and including the 64th pump, the expected value of pumping the balloon is positive; after 64, the expected value is (increasingly) negative. It is, therefore, optimal to aim for 64 pumps on every balloon, and then stop. Choosing to pump more *or* fewer than 64 times will decrease expected earnings; and the farther one deviates from the optimum, the lower the expected earnings become (Lejuez et al., 2002; Pleskac et al., 2008; Wallsten et al., 2005). Remarkably, in most trials, participants stop pumping the balloon far before the optimal stopping point (Lejuez et al., 2002). In fact, the average adjusted pump score is typically between 26 and 35 (Pleskac et al., 2008). Real-world risk-avoiders and risk-takers alike rarely pump the balloon enough times to maximize their expected earnings. This is less of a problem in the automatic BART, although participants there still pump fewer than 64 times on average. For example, two recent studies reported averages of 61.9 (Bernoster et al., 2019) and 58.5 pumps (De Groot & van Strien, 2019).

It is yet unknown exactly why participants often stop pumping before they reach the optimal point, but various factors may play a role. First, since the original BART requires participants to inflate balloons one pump at a time, it is plausible that they get tired of pumping after a while. Second, participants may want to limit their effort out of laziness or a desire to finish early (but see Young & McCoy, 2019). Third, they may become satiated: due to diminishing marginal returns, adding 5 cents to a growing temporary bank may stop being an attractive prospect well before reaching pump 64. Fourth, participants may need time to learn which strategy results in maximal earnings (Lejuez et al., 2002). This conjecture is supported by the observation that participants in both the original and the automatic BART on average press closer to the normative solution in the final block of 10 trials than they do in previous blocks (De Groot & van Strien, 2019; Lejuez et al., 2002). It also corresponds with the presumed shift from deciding under uncertainty to deciding under risk. In the BART, learning the optimal solution is hard, as the range of possible explosion points is large (1-128), and individual explosions provide limited feedback. This is in line with findings by Lejuez et al. (2002), who show that larger explosion ranges result in larger relative deviations from the optimum.

The fact that participants in the BART often stop pumping before the optimal stopping point has serious implications for how the data can be interpreted. Up to 64 pumps, the risk they take can be characterized as *adaptive* or *functional*, as it results in higher earnings. After that point, it can be considered *maladaptive* or *dysfunctional*, as it reduces expected earnings. Since people generally pump fewer than 64 times, the BART cannot properly differentiate between adaptive and maladaptive risk behavior, neither within nor between participants. A second, related problem is that experimental manipulations meant to increase risk-taking (such as adding time pressure or administering a certain drug) generally do not lead to lower earnings, as even the resulting higher pump numbers usually do not exceed 64 (Pleskac et al., 2008). For example, if a manipulation causes participants to take more risk and press 50 instead of 30 times, they are actually, on average, *better* off than before, the opposite of what one would expect in real life. In short, if participants mostly stay under 64 pumps, they simply never reach the point where taking more risk becomes disadvantageous, which limits the conclusions one can draw from the data.

The most straightforward way to mitigate these problems may be the modified BART developed by Pleskac et al. (2008), which differs from the original task in three ways. First, it involves an automatic response mode: participants indicate their intended number of pumps at the start of each trial, after which the balloon automatically inflates to the corresponding size (or until it bursts). Although meant to mitigate censoring, this adjustment may also prevent people from getting tired of pumping and from wanting to finish the task sooner. Second, the adjusted task provides explicit feedback about the explosion point of *every* balloon, not merely of those that actually explode. This may improve participants' learning across trials. Third, participants are (truthfully) informed that the range of pump numbers is 1-128 and that the best overall number of pumps is 64, further increasing the amount of information at their disposal.

These three modifications together successfully moved participants' behav-

Pump Number (A)	Balloon Value Before Pump (B)	Balloon Value After Pump (C)	Chance of Explosion (D)	Chance of Success (E)	Expected Value of Current Pump (F)	Expected Value of All Remaining Pumps (G)
1	€ -	€ 0.05	0.00781	0.99219	€ 0.05	€ 1.60
2	€ 0.05	€ 0.10	0.00787	0.99213	€ 0.05	€ 1.56
3	€ 0.10	€ 0.15	0.00794	0.99206	€ 0.05	€ 1.53
4	€ 0.15	€ 0.20	0.00800	0.99200	€ 0.05	€ 1.49
5	€ 0.20	€ 0.25	0.00806	0.99194	€ 0.05	€ 1.45
(...)						
62	€ 3.05	€ 3.10	0.01493	0.98507	€ 0.00	€ 0.00
63	€ 3.10	€ 3.15	0.01515	0.98485	€ 0.00	€ 0.00
64	€ 3.15	€ 3.20	0.01538	0.98462	€ 0.00	€ 0.00
65	€ 3.20	€ 3.25	0.01563	0.98438	€ 0.00	€ 0.00
66	€ 3.25	€ 3.30	0.01587	0.98413	€ 0.00	€ 0.00
(...)						
124	€ 6.15	€ 6.20	0.2	0.8	€ -1.19	€ -1.19
125	€ 6.20	€ 6.25	0.25	0.75	€ -1.51	€ -1.51
126	€ 6.25	€ 6.30	0.33333	0.66667	€ -2.05	€ -2.05
127	€ 6.30	€ 6.35	0.5	0.5	€ -3.13	€ -3.13
128	€ 6.35	€ 6.40	1	0	€ -6.35	€ -6.35

TABLE 1 Changing Balloon Values, Explosion and Success Chances, and Expected Values Across Balloon Pumps. *Note*: The expected value of the current pump (F) is computed by multiplying the success chance (E) by 0.05, then subtracting the product of the explosion chance (D) and the balloon value before the pump (B) [$F = E * 0.05 - D * B$]. Alternatively, one can also take into account the expected value of any subsequent pumps, insofar as they are advantageous (G). This results in somewhat different values, but an identical tipping point at 64.

ior closer to the normative solution of 64, with an average pump score of 57.7 for females and 63.7 for males (Pleskac et al., 2008). Part of this effect can be attributed to the automatic response mode, as these averages are higher than those from a manual BART with full feedback and strategy instructions added. Since this manual BART itself resulted in higher averages than the original BART, the feedback and instructions likely also contributed to the effect (Lejuez et al., 2002). Recent research, however, indicates that informing participants about the optimal strategy is not necessary, and even ill-advised. Two studies using an automatic BART with full feedback – but without strategy instructions – found equally high pump averages as did Pleskac and colleagues (Bernoster et al., 2019; De Groot & van Strien, 2019). Additionally, these studies found that a subgroup of participants – often from a STEM background – seem to infer the optimal strategy without any help.[2] Their repeated 64-answers, therefore, reflect cognitive ability rather than risk propensity and reduce task variability. Informing participants about the optimal strategy can increase such problematic responses. Therefore, it seems best to add automatic responses and full feedback to the BART, but not strategy instructions. This will likely elicit sufficiently high pump averages, without compromising the validity of the task.

Discussion

Since it was first published in 2002, the BART has become one of the most popular tools in psychology to gauge individuals' propensity for risk-taking. Halfway 2020, the original article describing the BART (Lejuez et al., 2002) had been cited over 1100 times in Scopus, most often in journals on decision research, addiction, and neuropsychology. This popularity is well-founded. The BART succeeds in recreating the 'natural' feeling of exhilaration and tension people experience when taking risk, and thus has excellent *ecological validity*. Furthermore, it correlates well with self-reported risk-related constructs, such as impulsivity and sensation-seeking, and with real-world risk behaviors, like polydrug use and unsafe sex, supporting its *convergent validity*. Lastly, it does not correlate with constructs like depression and anxiety, endorsing its *discriminant validity* (Lejuez et al., 2002). But despite these qualities, the BART suffers from methodological problems, most of which have been acknowledged in previous research as negatively impacting its rigor. The present paper is the first to give a comprehensive overview of these problems.

The *first* problem concerns the lack of clarity as to whether decisions in the BART are made under uncertainty (where outcome probabilities are unknown) or risk (where they are known). Since participants are not given any information about the explosion probabilities, they first decide under uncertainty, which then gradually shifts towards risk as they learn more about the probabilities in the task. As it is unclear exactly when this shift takes place, it is difficult to determine whether a given decision is made under uncertainty, risk, or something in between. The *second* problem concerns statistical censoring, which occurs in trials where the balloon explodes, as participants are then prevented from taking additional risk. As a result, the average number of times participants pump the balloon underestimates their risk propensity.

2. The relevant data collected by Bernoster et al., 2019 and De Groot and van Strien, 2019 on individual answering patterns was not published but can be shared upon request.

Third, the BART confounds risk with expected value. Since these constructs change simultaneously throughout a trial, participants' pump behavior again does not reflect risk propensity, as decisions are influenced by both risk and expected value. This also means that the task is poorly decomposable, as it cannot disentangle the motives underlying a pump decision. A *final* problem concerns the task's normative solution. In the majority of trials, participants stop pumping before the point where expected earnings are maximized. Therefore, participants mostly take adaptive risk, which leads to higher earnings. Maladaptive risk-taking hardly occurs, even though one would expect to see such behavior in certain cases.

Despite these problems, much of the research up to now has focused on the empirical findings produced by the BART, rather than on the task itself, with the majority of researchers using the task without critically reviewing whether its problems interfere with their aims. This can have undesirable consequences, such as when it leads to false positives or false negatives. For example, one may fail to show a relationship which only exists for decisions characterized by risk, as some trials in the BART are characterized by uncertainty instead. Conversely, a hypothesis may pertain to people's response to changing risk and be unjustly supported, as in the BART, risk and expected value simultaneously change and impact individuals' behavior. Finding *true* positives and negatives hinges on several factors, an important one being the validity of the measurement instrument. Any data resulting from instruments that suffer from methodological or interpretational problems is of limited value to understanding the concepts they are supposed to operationalize.

For these reasons, it is imperative that researchers critically evaluate the 'fit' between their research and the BART before deciding on using it. For many research aims, one will now see that the original BART does not suffice. Yet despite these 'burst beliefs', there are three types of approaches one can take to account for its limitations. *First*, data from the original BART can be analyzed using a different, more informative index than Lejuez et al.'s average adjusted pump score. For example, the models by Wallsten et al. (2005) break down behavior into risk-taking, response consistency, and learning. In addition, computational models can be used to take into account censoring and to provide an index of uncensored risk-taking in the BART (Dijkstra et al., 2020; Tobin, 1958; Weller et al., 2019). A *second* way of dealing with the BART's limitations is by modifying the task, for example by rigging it (Figner et al., 2009; Slovic, 1966), providing additional feedback, or automating the responses (Pleskac et al., 2008). *Third*, one may consider using a different task. This can be an existing (sequential) risk-taking task, like the Columbia Card Task (Figner et al., 2009), which performs better in terms of decomposability than the BART. Alternatively, researchers should consider creating a custom task that exactly suits their research, avoiding methodological flaws that could endanger the soundness of their conclusions. For instance, a task developed by Schmidt et al. (2013) involves decisions under conditions of explicit risk and does not confound risk with expected value. An important goal to keep in mind when designing such bespoke tasks is to combine strong ecological validity with methodological rigor (Schonberg et al., 2011).

Clearly, none of the solutions proposed can be considered a 'universal' fix that solves all of the BART's problems. Depending on the aims of any given study, certain problems will be more pressing than others, indicating the (type of) solutions most needed. By combining solutions, researchers could work towards a task that can operationalize risk propensity without substantial methodological or interpretational problems. For example, an automatic BART with full feedback and explicit information on the probability distribution provides uncensored decisions made under clear risk that are at times risky enough to be maladaptive. If the resulting data from this adapted BART are then analyzed using a model like that by Wallsten et al. (2005) or that by van van Ravenzwaaij et al. (2011), all problems reviewed in the current commentary would be addressed. However, this does not necessarily mean that this combination of solutions constitutes a universal fix after all, as the BART may face more problems than the ones discussed here. In all likelihood, the present review is not exhaustive. Researchers using the BART may know of additional problems, although this is unlikely to show in their work, as journals – and by extension researchers – do not consider 'failure' a popular publishing theme (Ferguson & Heene, 2012; Song et al., 2009). Therefore, it is important for researchers to not only critically evaluate the instruments they use but to disclose these evaluations as well, so that any and all methodological shortcomings can be openly discussed and addressed, improving the quality of the measures used.

Conclusion

The present paper is the first to review the methodological shortcomings of the Balloon Analogue Risk Task, a highly popular risk-taking task in psychology. The main problems identified are the ambiguity between uncertainty and risk, censoring of observations, confounding of risk and expected value, and poor decomposability into adaptive and maladaptive risk-taking. In addition, the paper reviews solutions that mitigate these problems. By presenting this first-time inventory, the paper highlights earlier mentions of problems in the BART as well as proposed solutions. It calls for a critical attitude towards the BART and experimental tasks in general, as their design deserves at least as much attention as the findings they produce. It also sets the agenda for testing and comparing different tasks and task versions, to explore which designs result in the best usability, reliability, and validity, so that risk propensity can be measured in the most accurate way possible.

References

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. https://doi.org/10.1016/0010-0277(94)90018-3

Bernoster, I., De Groot, K., Wieser, M. J., Thurik, R., & Franken, I. H. (2019). Birds of a feather flock together: Evidence of prominent correlations within but not between self-report, behavioral, and electrophysiological measures of impulsivity. *Biological Psychology*, *145*, 112–123. https://doi.org/10.1016/j.biopsycho.2019.04.008

Bishara, A. J., Pleskac, T. J., Fridberg, D. J., Yechiam, E., Lucas, J., Busemeyer, J. R., Finn, P. R., & Stout, J. C. (2009). Similar processes despite divergent behavior in two commonly used measures of risky decision making. *Journal of Behavioral Decision Making*, *22*, 435–454. https://doi.org/10.1002/bdm.641

Blais, A., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, *1*, 33–47. https://doi.org/10.1037/t13084-000

Brand, M., Fujiwara, E., Borsutzky, S., Kalbe, E., Kessler, J., & Markowitsch, H. J. (2005). Decision-making deficits of Korsakoff patients in a new gambling task with explicit rules: Associations with executive functions. *Neuropsychology*, *19*, 267–277. https://doi.org/10.1037/0894-4105.19.3.267

Brand, M., Labudda, K., & Markowitsch, H. J. (2006). Neuropsychological correlates of decision-making in ambiguous and risky situations. *Neural Networks*, *19*, 1266–1276. https://doi.org/10.1016/j.neunet.2006.03.001

Brand, M., Recknor, E. C., Grabenhorst, F., & Bechara, A. (2007). Decisions under ambiguity and decisions under risk: Correlations with executive functions and comparisons of two different gambling tasks with implicit and explicit rules. *Journal of Clinical and Experimental Neuropsychology*, *29*, 86–99. https://doi.org/10.1080/13803390500507196

Buelow, M. T., & Suhr, J. A. (2009). Construct validity of the iowa gambling task. *Neuropsychology Review*, *19*, 102–114. https://doi.org/10.1007/s11065-009-9083-4

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. https://doi.org/10.1037/1040-3590.14.3.253

Crosetto, P., & Filippin, A. (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, *47*, 31–65. https://doi.org/10.1007/s11166-013-9170-z

De Groot, K., & Thurik, R. (2018). Disentangling risk and uncertainty: When risk-taking measures are not about risk. *Frontiers in Psychology*, *9*, 21–94. https://doi.org/10.3389/fpsyg.2018.02194

De Groot, K., & van Strien, J. W. (2019). Event-related potentials in response to feedback following risk-taking in the hot version of the Columbia Card Task. *Psychophysiology*, *56*(9), e13390. https://doi.org/10.1111/psyp.13390 00000 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.13390

Dijkstra, N. F., Tiemeier, H., Figner, B. C., & Groenen, P. J. (2020). *A censored mixture model for modeling risk taking*.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*, 380–417. https://doi.org/10.1037/h0053870

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, *75*, 643–669. https://doi.org/10.2307/1884324

Euteneuer, F., Schaefer, F., Stuermer, R., Boucsein, W., Timmermann, L., Barbe, M. T., Ebersbach, G., Otto, J., Kessler, J., & Kalbe, E. (2009). Dissociation of decision-making under ambiguity and decision-making under risk in patients with Parkinson's disease: A neuropsychological and psychophysiological study. *Neuropsychologia*, *47*, 2882–2890. https://doi.org/10.1016/j.neuropsychologia.2009.06.014

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*, 555–561. https://doi.org/10.1177/1745691612459059

Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: Age differences in risk taking in the Columbia Card Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 709–730. https://doi.org/10.1037/a0014983

Grable, J., & Lytton, R. H. (1999). Financial risk tolerance revisited: The development of a risk assessment instrument. *Financial Services Review*, *8*, 163–181. https://doi.org/10.1016/S1057-0810(99)00041-4

Gu, R., Zhang, D., Luo, Y., Wang, H., & Broster, L. S. (2018). Predicting risk decisions in a modified balloon analogue risk task: Conventional and single-trial ERP analyses. *Cognitive, Affective, & Behavioral Neuroscience*, *18*, 99–116. https://doi.org/10.3758/s13415-017-0555-3

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description-experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, *21*, 493–518. https://doi.org/10.1002/bdm.598

Knight, F. H. (1921). *Risk, uncertainty and profit*. New York, NY, Sentry Press.

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2014). Individual differences in risky decision making: A meta-analysis of sensation seeking and impulsivity with the balloon analogue risk task. *Journal of Behavioral Decision Making*, *27*, 20–36. https://doi.org/10.1002/bdm.1784

Lauriola, M., & Weller, J. (2018). Personality and risk: Beyond daredevils – risk taking from a temperament perspective (E. L. Raue & B. Streicher, Eds.). In E. L. Raue & B. Streicher (Eds.), *M*. Psychological perspectives on risk and risk analysis, Springer.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The balloon analogue risk task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84. https://doi.org/10.1037/1076-898X.8.2.75

Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the Iowa Gambling Task. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 16075–16080. https://doi.org/10.1073/pnas.0406666101

Pleskac, T. J., Wallsten, T. S., Wang, P., & Lejuez, C. (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

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, *25*, 303–325. https://doi.org/10.1016/0304-4076(84)90004-6

Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, *23*, 1–14. https://doi.org/10.1002/bdm.681

Rogers, R. D., Owen, A. M., Middleton, H. C., Williams, E. J., Pickard, J. D., Sahakian, B. J., & Robbins, T. W. (1999). Choosing between small, likely

rewards and large, unlikely rewards activates inferior and orbital prefrontal cortex. *The Journal of Neuroscience*, *19*, 9029–9038. https://doi.org/10.1523/JNEUROSCI.19-20-09029.1999

Sattler, J. M., & Ryan, J. J. (2009). *Assessment with the WAIS-IV*. Jerome M. Sattler Publisher.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schmidt, B., Kessler, L., Holroyd, C. B., & Miltner, W. H. (2019). Wearing a bike helmet leads to less cognitive control, revealed by lower frontal midline theta power and risk indifference. *Psychophysiology*, *56*. https://doi.org/10.1111/psyp.13458

Schmidt, B., Mussel, P., & Hewig, J. (2013). I'm too calm – Let's take a risk! On the impact of state and trait arousal on risk taking. *Psychophysiology*, *50*, 498–503. https://doi.org/10.1111/psyp.12032

Schonberg, T., Fox, C. R., & Poldrack, R. A. (2011). Mind the gap: Bridging economic and naturalistic risk-taking with cognitive neuroscience. *Trends in Cognitive Sciences*, *15*, 11–19. https://doi.org/10.1016/j.tics.2010.10.002

Shuttleworth, M. (2008). *Operationalization*. https://explorable.com/operationalization

Slovic, P. (1966). Risk-taking in children: Age and sex differences. *Child Development*, *37*, 169–176. https://doi.org/10.2307/1126437

Song, F., Parekh-Bhurke, S., Hooper, L., Loke, Y. K., Ryder, J. J., Sutton, A. J., Hing, C. B., & Harvey, I. (2009). Extent of publication bias in different categories of research cohorts: A meta-analysis of empirical studies. *BMC Medical Research Methodology*, *9*, 79. https://doi.org/10.1186/1471-2288-9-79

Starcke, K., Tuschen-Caffier, B., Markowitsch, H. J., & Brand, M. (2009). Skin conductance responses during decisions in ambiguous and risky situations in Obsessive-Compulsive Disorder. *Cognitive Neuropsychiatry*, *14*, 199–216. https://doi.org/10.1080/13546800902996831

Starcke, K., Tuschen-Caffier, B., Markowitsch, H. J., & Brand, M. (2010). Dissociation of decisions in ambiguous and risky situations in Obsessive-Compulsive Disorder. *Psychiatry Research*, *175*, 114–120. https://doi.org/10.1016/j.psychres.2008.10.022

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*, 24–36. https://doi.org/10.2307/1907382

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E. (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

Volz, K. G., & Gigerenzer, G. (2012). Cognitive processes in decisions under risk are not the same as in decisions under uncertainty. *Frontiers in Neuroscience*, *6*, 105. https://doi.org/10.3389/fnins.2012.00105

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

Weller, J. A., King, M. L., Figner, B., & Denburg, N. L. (2019). Information use in risky decision making: Do age differences depend on affective context? *Psychology and Aging*, *34*, 1005–1020. https://doi.org/10.1037/pag0000397

Young, M. E., & McCoy, A. W. (2019). Variations on the balloon analogue risk task: A censored regression analysis. *Behavior Research Methods*, *51*, 2509–2521. https://doi.org/10.3758/s13428-018-1094-8

License