

Representative Design in Psychological Assessment: A Case Study Using the Balloon Analogue Risk Task (BART)

Markus D. Steiner and Renato Frey

Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel

Representative design refers to the idea that experimental stimuli should be sampled or designed such that they represent the environments to which measured constructs are supposed to generalize. In this article we investigate the role of representative design in achieving valid and reliable psychological assessments, by focusing on a widely used behavioral measure of risk taking—the Balloon Analogue Risk Task (BART). Specifically, we demonstrate that the typical implementation of this task violates the principle of representative design, thus conflicting with the expectations people likely form from real balloons. This observation may provide an explanation for the previously observed limitations in some of the BART's psychometric properties (e.g., convergent validity with other measures of risk taking). To experimentally test the effects of improved representative designs, we conducted two extensive empirical studies ($N = 772$ and $N = 632$), finding that participants acquired more accurate beliefs about the optimal behavior in the BART due to these task adaptations. Yet, improving the task's representativeness proved to be insufficient to enhance the BART's psychometric properties. It follows that for the development of valid behavioral measurement instruments—as are needed, for instance, in functional neuroimaging studies—our field has to overcome the philosophy of the “repair program” (i.e., fixing existing tasks). Instead, we suggest that the development of valid task designs requires novel ecological assessments, aimed at identifying those real-life behaviors and associated psychological processes that lab tasks are supposed to capture and generalize to.



Keywords: representative design, BART, risk taking

Supplemental materials: <https://doi.org/10.1037/xge0001036.supp>

Various psychological assessments are routinely performed by means of behavioral tasks, including the measurement and modeling of individual differences in risk taking (Frey et al., 2017, 2020; Lauriola et al., 2014; Lejuez, Aklin, Jones, et al., 2003; Mishra & Lalumière, 2011; Tisdall et al., 2020). Although such task-based assessments of *revealed preferences* have been considered the gold standard in some fields of psychology and economics (e.g., Beshears et al., 2008; Charness et al., 2013), recent evidence has highlighted substantial psychometric limitations of this measure-

ment approach (e.g., Beauchamp et al., 2017; Berg et al., 2005; Eisenberg et al., 2019; Frey et al., 2017; Lönnqvist et al., 2015; Millroth et al., 2020). Valid and reliable alternatives do exist in the form of self-report measures (e.g., Arslan et al., 2020; Frey et al., 2017; Steiner et al., *in Press*), yet behavioral tasks may continue to be indispensable for certain applications, such as in research on the functional neural architecture of risk taking, which typically rests on the simulation of risk-taking behaviors in the fMRI scanner (e.g., Helfinstein et al., 2014; Li et al., 2019; Rao et al., 2008; Schonberg et al., 2011; Tisdall et al., 2020). Moreover, incorporating both revealed and stated preferences in a multimethod approach may prove beneficial for understanding and predicting real-life behavior (e.g., Lejuez et al., 2002; Sharma et al., 2014; Wallsten et al., 2005).

In this article, we build on an argument originally put forth by Brunswik and examine the role of *representative design* (Brunswik, 1956; Gibson, 1986; Hammond, 1966; Stoffregen et al., 2003; for an overview see Araújo et al., 2007 and Dhimi et al., 2004) in behavioral measures of risk taking. Representative design (not to be confused with *ecological validity*; Araújo et al., 2007) refers to the idea that experimental stimuli should be sampled or designed such that

 Markus D. Steiner,  Renato Frey. This work was supported by a grant of the Swiss National Science Foundation (PZ00P1_174042) provided to Renato Frey. We thank Laura Wiles for proofreading, and Alexandra Bagaïni, Silvia Grieder, and the members of the Center for Cognitive and Decision Sciences for helpful comments. Data, analysis code, screenshots of the experimental paradigm, and the preregistration are available at <https://osf.io/kxp8t>.

Corresponding author: Markus D. Steiner, Department of Psychology, University of Basel, Missionsstrasse 60-62, 4055 Basel, Switzerland. E-mail: markus.steiner@unibas.ch.

they adequately *represent* the environments to which measured constructs are supposed to generalize, and that “experimenters should avoid oversampling highly improbable [...] variables in the intended behavioral setting” (Araújo et al., 2007, p. 73). Specifically, we argue that violations of representative design may contribute to the poor psychometric properties of behavioral risk-taking measures as have been observed in previous research, such as low convergent validity or low test–retest reliability (Beauchamp et al., 2017; Berg et al., 2005; Eisenberg et al., 2019; Frey et al., 2017; Lönnqvist et al., 2015; Mata et al., 2018; Slovic, 1962)—and thus ultimately hamper a successful assessment of meaningful individual differences. This article illustrates this argument, and systematically examines the potential benefits of using improved representative designs, by focusing on the Balloon Analogue Risk Task (BART).

The BART: A Prominent Behavioral Measure of Risk Taking

The BART is one of the most prominent behavioral measures used to gauge individual differences in risk taking, often employed in behavioral decision research (e.g., Lauriola et al., 2014; Lejuez et al., 2002; Wallsten et al., 2005), in clinical settings (e.g., Bornovalova et al., 2005; Hopko et al., 2006; Hunt et al., 2005), as well as in applied contexts (e.g., Aklin et al., 2005; Lejuez, Aklin, Zvolensky, & Pedulla, 2003). For instance, the BART has been used to predict interindividual differences in substance use (e.g., Campbell et al., 2013; Hanson et al., 2014; Hopko et al., 2006; Lejuez, Aklin, Jones, et al., 2003), to study the neural architecture of risk-taking behaviors in imaging studies (e.g., Helfinstein et al., 2014; Li et al., 2019; Rao et al., 2008; Tisdall et al., 2020), and to examine the genetic underpinnings thereof (Mata et al., 2012).

When completing the BART, participants sequentially inflate virtual balloons (typically 30) on a computer screen, earning a fixed amount of money for each successful inflation. If a balloon explodes, the money accrued in the current trial is lost. Participants are free to stop inflating a balloon at any time, to thus transfer their current gain to a safe account. At the onset of the task, participants are only told the amount of money they will earn for each successful inflation, that they will lose the money accrued in the current trial if the balloon bursts, as well as that at most the balloons can get as large as the whole screen. As such, participants initially face a situation of decisions under *uncertainty* (see Knight, 1921; Mousavi & Gigerenzer, 2014), because the risk of an explosion at different inflation stages remains unknown. With increasing experience, the task gradually transforms into a situation of decisions under *risk* (Knight, 1921; Mousavi & Gigerenzer, 2014), as the explosion probabilities can in principle be learned—at least approximatively.

The BART is attractive as it resembles many real-life de-

cision problems in at least three key aspects: On the one hand, it mirrors the fact that in many risky situations not all stochastic properties are known a priori but have to be learned through experience (e.g., Frey, 2020; Frey et al., 2015; Hertwig et al., 2004). On the other hand, the sequential nature of the BART creates a “sense of escalating tension and exhilaration” (Schonberg et al., 2011, p. 16), mimicking the thrill that individuals may feel in many risk-taking decisions in real life (e.g., whether to stay invested in stocks before a looming stock market crash). Moreover, risk and reward are correlated in the BART, as they are in many real-life decisions involving risk and uncertainty (Pleskac & Hertwig, 2014; Pleskac et al., 2020).

In light of these attractive features, it may be somewhat surprising that several studies documented a relatively low convergent validity of the BART with measures tapping various constructs related to risk taking. For instance, one study found a maximum correlation of $r = .16$ between the BART and any of 38 multi-dimensional risk-taking measures, spanning indicators of domain-general and domain-specific risk preference, sensation seeking, impulsivity, and concrete real-life behaviors, as well as comprising different assessment methods (i.e., self-reported propensity measures, behavioral measures, and frequency measures; Frey et al., 2017). Alike, meta-analyses on the BART’s convergent validity reported similarly low correlations (i.e., $r = .14$ for sensation seeking, and $r = .10$ for impulsivity; Duckworth & Kern, 2011; Lauriola et al., 2014). Moreover, although multiple studies found associations of the BART with real-life behaviors (e.g., Aklin et al., 2005; Lejuez, Aklin, Jones, et al., 2003; Lejuez et al., 2007; Skeel et al., 2008), this has not consistently been the case (e.g., Frey et al., 2017; Hopko et al., 2006; Hunt et al., 2005; Lauriola et al., 2014; Schürmann et al., 2018)—and to date no meta-analysis exists yet to conclusively clarify this issue. Finally, although the BART exhibits a high test–retest reliability, especially in comparison with other behavioral tasks (Frey et al., 2017; White et al., 2008), it is somewhat lower as compared to respective self-report measures (e.g.; Frey et al., 2017; Mata et al., 2018). The question thus arises: What obstacles hinder the BART from capturing individual differences in risk taking more consistently, and how could such limitations potentially be fixed?

Challenges in the BART’s task design

Previous research concerning the BART’s task design has mainly revolved around two potential issues. First, it has been argued that learning may be difficult due to the asymmetric feedback provided (Pleskac et al., 2008). Removing learning requirements (i.e., either by informing participants upfront about the optimal number of inflations; or by implementing a related task that retains the BART’s basic structure yet has no learning demands) resulted in similar and partly stronger associations with some real-life behav-

iors (i.e., polydrug use; Pleskac, 2008; Pleskac et al., 2008). That is, whether or not the BART's learning requirement is ultimately a useful property may also depend on the particular real-life behaviors that are to be predicted (e.g., the extent to which these are decisions under uncertainty that involve a learning component).

Second, there has been a debate concerning people's representations of explosion probabilities in the BART: Early work relying on cognitive modeling concluded that participants may form an incorrect representation of the task's stochastic structure by assuming that explosion probabilities remain stationary across the sequential inflation process (Pleskac, 2008; Wallsten et al., 2005). However, more recent research, which has directly prompted participants to rate the probability that a balloon explodes at different inflation stages, has challenged this conclusion: According to participants' explicit ratings, they indeed expected a strong increase in the explosion probabilities during the sequential inflation process (Schürmann et al., 2018).

Here we would like to draw attention to yet another and independent, but potentially very fundamental issue in the BART's task design. Specifically, in order to trigger a sense of increasing tension during the sequential inflation process—as outlined above, an attractive feature that mimics many real-life situations—the conditional probability that a balloon explodes at inflation i (i.e., given that it has not exploded in the preceding $i - 1$ inflations; see the escalating purple curve in Figure 1a) is defined as

$$p(\text{expl}_i | \neg \text{expl}_{i-1}) = 1/(C - i + 1) \quad (1)$$

where C denotes the maximum capacity of the balloons, for example $C = 128$ (Lejuez et al., 2002).¹ Importantly, and as can be seen from the flat purple curve in Figure 1b, this stochastic structure results in a *uniform distribution* of explosion points. That is, when inflating all balloons to their explosion points, in the long run there will be the same number of explosions at every possible inflation stage (i.e., $p(\text{expl}_i) = 1/C$ for all inflation stages $i \in \{1, 2, \dots, C\}$).

Evidently, the typical implementation of the BART—hereinafter referred to as the BART_{uniform}—is in stark contrast to the stochastic structure to be expected from real balloons: Balloons of the same type can be expected to burst around one specific inflation stage, thus resulting in a distribution of explosions with a central tendency. To put this assumption to a simple test, we inflated 100 real balloons until they exploded, using a regular bicycle pump, and keeping record of the number of inflations. As to be expected, the resulting distribution of explosions (Figure 2) was much more aligned with a normal rather than a uniform distribution.² Hence, what are the potential consequences if representative design is violated in a behavioral task such as the BART?

Second, over-learning one's prior expectations may be es-

pecially challenging in the case of the BART_{uniform} because participants experience highly variable feedback—precisely due to the uniform distribution of explosions, which yields very early as well as very late explosions with the same likelihood. Furthermore, the highly variable explosion points may also lead to problematic order effects: Previous research has found a systematic influence of whether participants experience early or late explosions during the initial trials—requiring the order of explosions to be fixed across participants (Schürmann et al., 2018; Walasek et al., 2014). Thus, this second issue likely aggravates the consequences of the first issue.

Three issues associated with the lack of representative design in the BART

To date, the degree of representative design and its respective effects remain rarely tested for specific tasks, particularly in the context of psychological assessment. Although some studies have found mixed evidence concerning whether representative design and systematic design (i.e., the attempt to systematically design stimuli to be able to have maximal control over experimental manipulations) generally lead to substantially different effects (Dhimi et al., 2004), in the case of the BART one can conceive of at least three major issues.

First, and particularly in a “naturalistic” task such as the BART, participants do not start off as *tabula rasa* but with some prior beliefs (see also, Pleskac, 2008; Wallsten et al., 2005): Virtually everyone has inflated real balloons and acquired the expectation that explosions do not occur in an entirely unpredictable way—as is the case in the BART_{uniform}. Thus, in the process of turning this task from a situation of decisions under uncertainty into one of decisions under risk, participants may aim to learn (implicitly or explicitly) about several statistical properties, such as: “Around which value do most of the balloons explode?” In fact, due to the linear reward structure the expected payoffs are maximized when inflating all balloons to half of the maximum capacity (Figure 1d); and as this reward structure is transparent (i.e., participants know upfront that payoffs increase linearly with each inflation; Lejuez et al., 2002), the goal of maximizing payoffs reduces entirely to learning about the (mean of the) distribution of explosion points. Hence, the respective need to over-learn one's prior expectations about the functional form of the distribution of explosions may introduce undesirable noise in the BART_{uniform}, and may thus lead to distorted task representations—which could limit not only the task's test-retest reliability but also its convergent validity with related measures of risk taking.

¹If only one type of balloon is employed in an experiment, all balloons have, in principle, the same maximum capacity.

²Note that this distribution had somewhat fat tails and some degree of skewness, both of which may be related to the relatively small sample size of this brief experiment.

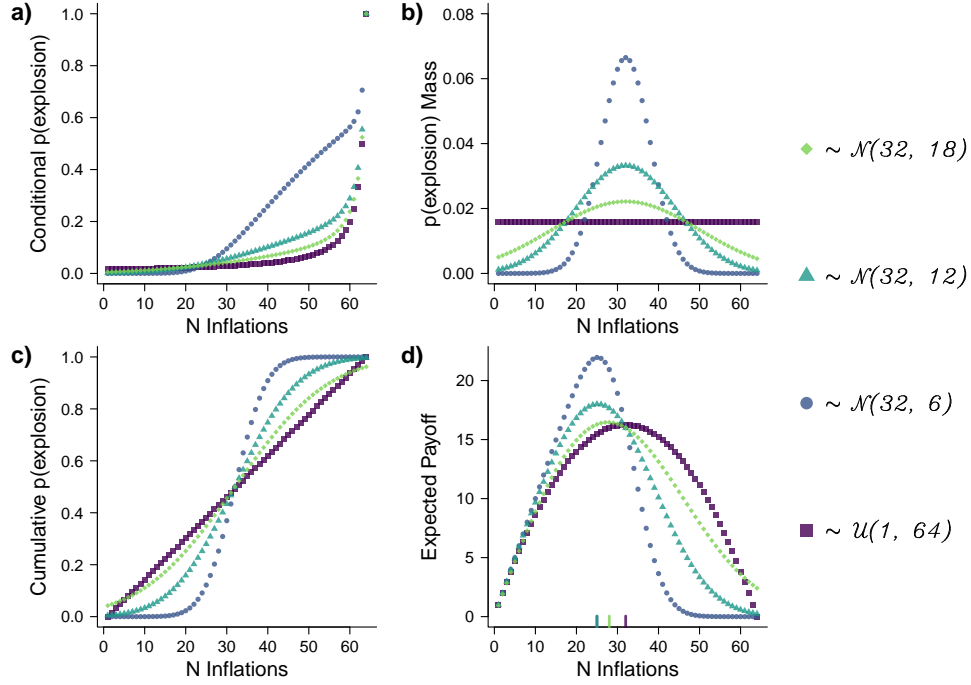


Figure 1

Illustration of four different task designs, each implementing a different stochastic structure in the BART. Colors/shapes indicate different distributions of explosion points, with purple/squared dots depicting the standard implementation of the BART (i.e., uniform distribution) and the other colors/shapes depicting more representative designs thereof (i.e., normal distributions). Panel a) shows the conditional explosion probabilities. Panel b) shows the probability masses of the explosion points. Panel c) shows the cumulative explosion probabilities. Panel d) shows the expected payoffs across inflation stages, and the colored ticks on the x-axis show the stage that maximizes the expected value, namely, 32, 28, 25, and 25 when explosion points are distributed as $\mathcal{U}(1, 64)$ as in the $\text{BART}_{\text{uniform}}$, $\mathcal{N}(32, 18)$ as in the $\text{BART}_{\text{normal-H}}$, $\mathcal{N}(32, 12)$ as in the $\text{BART}_{\text{normal-M}}$, and $\mathcal{N}(32, 6)$ as in the $\text{BART}_{\text{normal-L}}$.

Third, the payoff-maximizing behavior in the $\text{BART}_{\text{uniform}}$ consists of inflating the balloons up to the mean breaking point. Yet, around the mean breaking point there is no specific signal a participant could detect and exploit across trials—unlike in a distribution with a central tendency of explosion points (e.g., a normal distribution), where participants would spontaneously observe that relatively more balloons explode around a specific inflation stage. Thus, in order to adopt the objectively optimal behavior in the $\text{BART}_{\text{uniform}}$, participants have to obtain an estimate of C , as the mean of a uniform distribution (with a lower bound of 0) is defined as $\mu = \frac{C}{2}$. An estimate of C may be obtained through sequential updating of one's prior assumption of the balloons' maximum capacity (Wallsten et al., 2005), but this process is difficult due to the relatively few trials typically completed in the BART, as well as due to the asymmetric feedback provided. As a result, unless participants commit to a large number of purely exploratory trials, their estimates of C may be systematically biased downwards simply due to the particular task structure (Pleskac et al., 2008). Con-

sequently, even individuals who differ in their willingness to take risks may show very similar behaviors, which may lead to attenuated correlations with other measures of risk taking.

Taken together, these three issues may provide explanations for the reviewed limitations in the BART's psychometric properties. Thus, in what follows we will first report a reanalysis of five datasets, aimed at exploring the empirical evidence concerning whether participants' prior expectations indeed diverge from the distribution of explosion points as implemented in the $\text{BART}_{\text{uniform}}$. Then, we will report two empirical studies that systematically tested whether an improved representative design in the BART leads to an enhanced assessment of individual differences, which (a) may increase the convergent validity of the BART with measures of various constructs related to risk taking and (b) potentially boosts the BART's test-retest reliability.

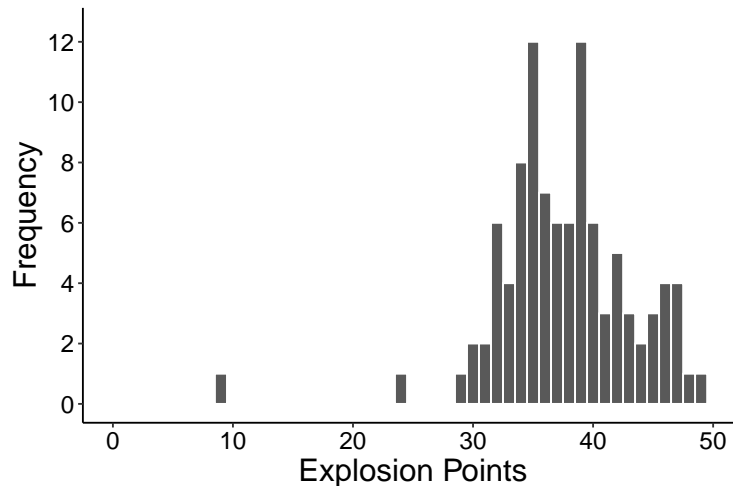


Figure 2

Distributions of explosion points of 100 real balloons, inflated with a bicycle pump.

Reanalysis of Five Datasets: People's Representations of the BART's Stochastic Structure

To explore people's expectations and representations concerning the stochastic structure of the BART, we first report a reanalysis of a number of datasets that comprise explicit judgments of explosion probabilities. Specifically, our reanalysis involves five datasets stemming from three studies: Frey et al., 2017 collected data from 1507 participants, Schürmann et al., 2018 collected data from 100 participants in study 1 and from 90 participants in study 2, and Steiner and Frey (2020) collected data from 31 participants.³ At the end of each of these studies (i.e., after having completed all 30 trials of the BART), participants were shown balloons inflated to different stages and were asked: "What do you think is the probability that the balloon will explode with one additional pump, given it is already inflated at this size?" (Schürmann et al., 2018, p. 4). Participants then provided their probability rating on a scale from 0% to 100%. In study 2 of Schürmann et al., 2018, participants provided these probability ratings twice, once after the first trial—thus also permitting some insights concerning participants' prior expectations—and once at the end of the task.

Schürmann et al., 2018 fitted psychometric functions to participants' ratings and visual inspection of the results (their Figures 3 and 5) suggests two conclusions: First, participants might generally have reported their beliefs that a balloon can be inflated *up to* different stages (i.e., cumulative probabilities; see the curves depicted in Figure 1c) rather than their beliefs that a balloon will explode at the next stage (i.e., conditional probabilities; see the curves depicted in Figure 1a; for similar findings, see Haffke & Hübner, 2019). Second and more importantly from the perspective of representative de-

sign, the shapes of the fitted psychometric functions suggest that participants may indeed have acquired the representation that the explosion points are normally and not uniformly distributed: In the case of a uniform distribution, cumulative probabilities would result in a linear function, whereas in the case of a normal distribution cumulative probabilities would result in a sigmoid function (see Figure 1c). The results of Schürmann et al., 2018 appear to be in line with the latter.

Method

To formally test these hypotheses, we fitted cumulative density functions (CDF) and conditional probability functions (CPF) of both a normal distribution and a uniform distribution to participants' probability ratings, and examined which function best described the data according to the least-squares criterion. For the two CDFs, we estimated two free parameters (i.e., mean and standard deviation in the case of the normal distribution, and the lower and upper bound in the case of the uniform distribution). For the two CPFs, we estimated two free parameters in the case of the normal distribution (i.e., the mean and standard deviation), and one free parameter in the case of the uniform distribution (i.e., the lower bound). Both CPFs used the maximum balloon capacity C as fixed upper bound, which was 128 in Schürmann et al., 2018 and in Frey et al., 2017, and 64 in Steiner and Frey, 2020.

Results

Figure 3 depicts the results of our reanalysis. In all datasets, the ratings of most participants were best described

³This dataset stems from a pilot study of a manuscript in preparation, see <https://osf.io/kxp8t> for the respective data and materials.

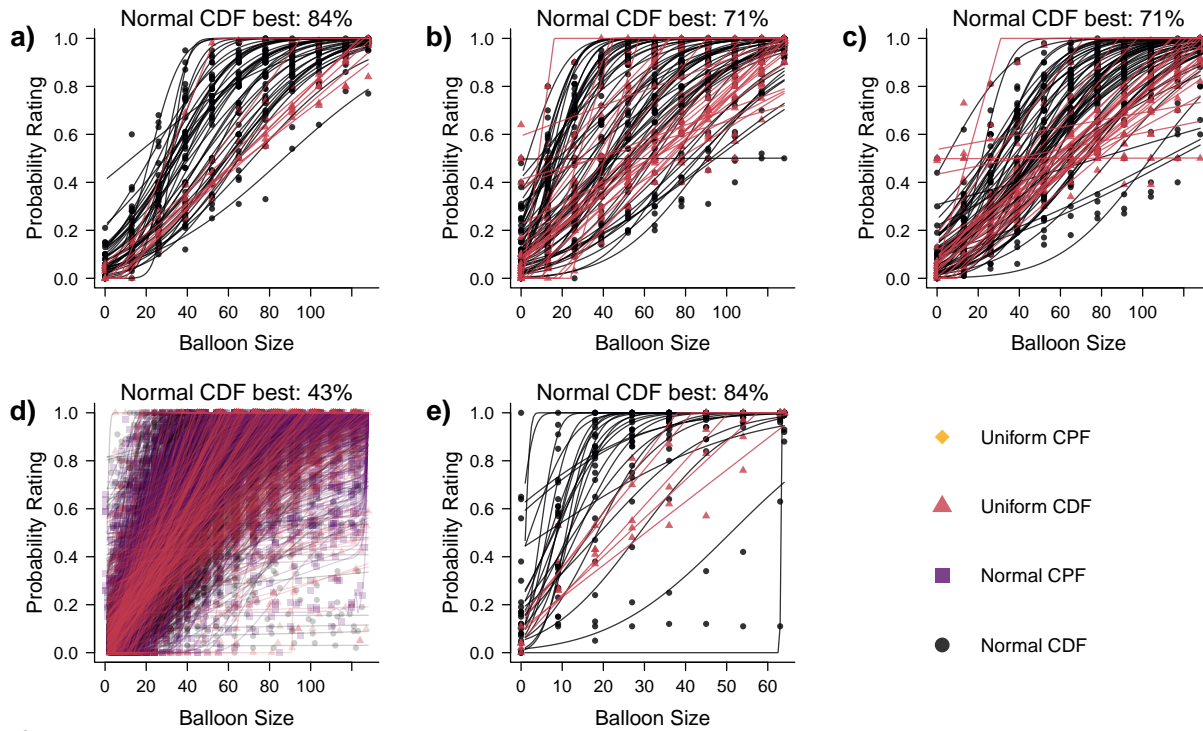


Figure 3

Reanalysis of the data from probability rating tasks. Panel a) shows the result of the reanalysis of study 1 from Schürmann et al., 2018. Panels b) and c) show the results of the reanalysis of study 2 from Schürmann et al., 2018 after participants have played one trial, and all 30 trials, respectively. Panel d) shows the results of the reanalysis of the data from Frey et al., 2017. Finally, panel e) shows the results of the reanalysis of the data from Steiner and Frey, 2020. Points represent the actual probability ratings. Lines are the predictions made by the models that best represent the respective participants. The line and point colors indicate the best fitting model. CDF = Cumulative density function. CPF = Conditional probability function.

by normal distributions. Specifically, the ratings of 84%, 71% and 71% of the participants in the different datasets from Schürmann et al., 2018, and 84% of the participants from Steiner and Frey, 2020, were best described by CDFs of normal distributions (i.e., the ratings of no participant were best described by a CPF). In the dataset of Frey et al., 2017 the ratings of 76% of participants were best described by a normal distribution (43% of participants by CDFs of normal distributions, and 33% of participants by CPFs of normal distributions). The ratings of the remaining 24% of participants were best described by a CDF of a uniform distribution (for a potential explanation of the somewhat different pattern in the latter dataset, see [online Supplemental Material Section 1](#)).

Discussion

Our reanalysis of five datasets consistently indicated that participants clearly exhibit a task representation that conflicts with the distribution of explosion points implemented in the $BART_{uniform}$: Most participants expected a normal distribu-

tion of explosion points—evidently the state of affairs in the real world (see Figure 2)—both in the beginning of the task (as assessed in Schürmann et al., 2018), and even after 30 trials of learning opportunity (as assessed in all four datasets).

Study 1: Does a More Representative Design Boost the BART's Convergent Validity With Other Measures of Risk Taking?

The goal of study 1 was to empirically test whether enhancing representative design in the BART improves the task's psychometric properties, thus permitting an improved assessment of participants' willingness to take risks. There exist multiple ways of implementing representative design: According to the definition of Hammond (1966) our brief test of how the explosions of real balloons are distributed (Figure 2) falls into the category of *substantive sampling*. Specifically, we have sampled real stimuli from the model behavior the experimental task was abstracted from. Naturally, such a direct implementation of representative design—which mirrors Brunswik's initial conception (Brunswik, 1956; see also

Dhami et al., 2004)—is not feasible or even desired in most assessment contexts (e.g., in online research). There is, however, another form of implementing representative design: *formal sampling*. It implies that formal, statistical properties of a judgment task are considered in the experimental design (Hammond, 1966). We followed this logic in study 1 by implementing the BART with (different types of) normal distributions of explosion points (i.e., BART_{normal}). We expected this change in the task architecture to lead to several improvements:

First, because explosion points are clustered in the BART_{normal}, participants experience more consistent feedback across trials. This should facilitate the acquisition of an appropriate task representation, particularly if participants expect (and thus aim to identify) such a clustering. Moreover, the more consistent feedback as compared to in the BART_{uniform} should in principle avoid the problem of systematic order effects in learning, potentially rendering the use of a fixed order of explosion points obsolete.

Second, unlike in the BART_{uniform}, in the case of a normal distribution there is no longer a need for an accurate estimate of the balloons' maximum capacity C to gauge the objectively optimal behavior. Instead, as an approximation of the number of inflations that maximizes payoffs, one can directly learn about the average explosion point (Figure 1d). This may be a more natural process, as learning about the mean of the balloons' explosion points can occur directly due to a noticeable increase in the the number of balloons that explode around a specific inflation stage. Although the asymmetric feedback in the BART may still lead to an underestimation of the number of inflations that maximizes payoffs, this should occur substantially less so than in the BART_{uniform}.

Third, we expected that these improvements would ultimately lead to an improved assessment of individual differences: On the one hand and as can be seen in Figure 1a, the conditional probabilities of an explosion may still create a desirable sense of escalating tension and exhilaration in the BART_{normal}. On the other hand, participants' overt risk-taking behavior may be a more direct expression of their willingness to take risks, due to reduced interindividual differences in participants' task representations. Hence, the convergent validity between the adjusted BART scores and other measures of risk taking should increase.

In our comparison of the BART_{normal} with the BART_{uniform}, we implemented three different normal distributions that had the same means but varied in terms of their standard deviation (i.e., we sampled a range of plausible learning environments for the BART_{normal}). Narrower distributions should lead to a reduced variability in participants' task representations. Yet, in the extreme case, a too narrow distribution may result in a trivial task, thus failing to capture any meaningful individual differences in risk taking. The distributions all had the same mean of

32, as in our implementation of the BART_{uniform} (which had a maximum capacity of 64). Moreover, the standard deviation of the widest normal distribution was explicitly chosen to match the standard deviation of the BART_{uniform}. To summarize, in study 1 we tested the following four hypotheses:

Hypothesis 1: General task representation: At the end of the task, participants believe that the explosion points cluster around a mean value rather than being uniformly distributed, irrespective of the actual distributional form implemented (i.e., BART_{uniform} vs. BART_{normal}). Moreover, within the BART_{normal}, we expected this belief to be increasingly stronger, the smaller the standard deviations of the distributions become.

Hypothesis 2: Beliefs about optimal behavior: At the end of the task, participants' beliefs about the inflation stage that maximizes their payoffs exhibit less variability between participants in the BART_{normal} as opposed to in the BART_{uniform}. Moreover, we expected these beliefs to be closer to the value that actually maximizes payoffs in the former as compared to in the latter.

Hypothesis 3: Overt risk-taking behavior: On average, participants' adjusted BART scores are closer to the optimal value and exhibit less variability between participants in the BART_{normal} than in the BART_{uniform}. Within the BART_{normal}, we expected that the adjusted BART scores are increasingly closer to the optimal value and exhibit less variability, the smaller the standard deviations of the distributions become.

Hypothesis 4: Convergent validity: As the distribution of adjusted BART scores in the BART_{normal} potentially reflects individual differences in participants' willingness to take risks more directly, we expected a higher convergent validity between adjusted BART scores and various other measures of risk taking (i.e., propensity and frequency measures) in the BART_{normal} as compared to in the BART_{uniform}.

Methods

Both empirical studies of this article were preregistered on the Open Science Framework. The preregistration, data files, and analysis scripts can be accessed via <https://osf.io/kxp8t>. Both empirical studies were approved by the local institutional review board (Number 020-19-1).

Participants and sample characteristics

Based on an a priori power analysis (see preregistration) we collected data of 800 participants on Amazon Mechanical Turk (MTurk). We imposed the following inclusion criteria: based in the United States, at least 18 years old, at least 500 completed tasks (HITs) on MTurk, and an acceptance rate of at least 99%. Moreover, only data were included of participants who passed at least one out of two attention check questions (see preregistration), who provided a rating of at least 25 on a scale from 0 to 100 concerning how focused they were during the study, and who confirmed to have completed the study on a desktop computer or a laptop. Of these 800 participants, the data of 28 contained missing values and we used list-wise deletion of these data, resulting in a final dataset consisting of data from 772 participants (47.8% female; $M_{age} = 38.0$, $SD_{age} = 11.1$; highest completed degree: 0.8% no high school, 37.1% high school, 40.3% bachelor, 10.1% master, 10.5% professional, 1.3% doctor; job status: 3.5% student, 11.0% unemployed, 82.5% working, 3.0% retired). On average, study completion took 13 minutes. Participants were reimbursed with a fixed payment of 10 cents and a performance contingent bonus payment, resulting in an average reimbursement of 4.36 USD.⁴

Materials and procedure

The whole study was conducted online on participants' own devices. After providing informed consent, participants completed the BART in one of four randomly assigned between-subjects conditions (see next paragraph). Upon completion of the BART, participants provided their beliefs about (a) the form of the underlying distribution (clustered explosion points vs. uniformly distributed explosion points; using a slider ranging from 0 to 50)—a procedure that we slightly revised and reimplemented in study 2—and (b) the optimal behavior in the BART (in randomized order). Then, participants completed (in a randomized order) the General Risk Propensity Scale (GRiPS; Zhang et al., 2018), the general and domain-specific risk items used in the German Socioeconomic Panel (SOEP; e.g., Dohmen et al., 2011, i.e., *propensity measures* in which participants self-report their risk preferences), and an assessment of real-life risk-taking behavior in different domains (i.e., *frequency measures*, in which participants report the frequency with which they engaged in different risky behaviors within the last year). Finally, participants reported their age, sex, job status, and highest education; how focused they were during the study, and the device they used to complete the study; and were given the possibility to provide free-text feedback. Screenshots of the study are provided at <https://osf.io/kxp8t>.

BART. Each participant was randomly assigned to one of the four between-subjects conditions; namely BART_{uniform} (N = 190), BART_{normal-H} (N = 195), BART_{normal-M} (N = 197),

and BART_{normal-L} (N = 190). In the BART_{uniform} the balloons' explosion points were drawn from $\mathcal{U}(1, 64)$. In the three versions of the BART_{normal}, the explosion points were drawn from three different normal distributions that varied in terms of their standard deviation (SD); namely, $\mathcal{N}(32, 12)$ representing a high SD (BART_{normal-H}), $\mathcal{N}(32, 18)$ representing a medium SD (BART_{normal-M}), and $\mathcal{N}(32, 6)$ representing a low SD (BART_{normal-L}). In all four implementations, balloons had a maximum capacity C of 64. Participants earned 1 cent per successful inflation; that is, their bonus equalled the sum of the number of inflations of balloons that did not explode.

Some of the previous research relying on the BART_{uniform} implemented a predefined sequence of explosions, in order to avoid random variation of samples and thus to reduce the risk of order effects across participants (Lejuez et al., 2002; Schürmann et al., 2018; Walasek et al., 2014). Although in principle this should be less of a concern in the BART_{normal} (particularly in the implementation with small standard deviations), for reasons of comparability we also generated a fixed sequence of 30 explosion points, for each of the four conditions, that closely represented the underlying distribution (see Figure S1; the respective R script can be accessed via <https://osf.io/kxp8t>). The explosion points were ordered quasi-randomly to generate a fixed sequence of 30 trials, such that the first three balloons had explosion points larger than ten and smaller than 54, and such that in the first ten, the second ten, and the third ten balloons the following properties held: five explosion points were greater or equal to the mean and five were smaller or equal to the mean; the mean was within 32 ± 0.25 (see also, Lejuez et al., 2002, for a similar approach to balancing the distributions).

As main dependent variable of participants' behavior, we focused on the adjusted BART score that reflects the mean number of inflations across balloons that did not explode (Lejuez et al., 2002). Although the adjusted BART score is typically highly correlated with the BART score (i.e., the mean number of inflations across all balloons), it is routinely used in studies on the BART as it may better reflect participants' intended behavior (Lejuez et al., 2002; but see, Pleskac et al., 2008). Another dependent variable consists of the total number of explosions per participant. It has been argued that the latter is advantageous as compared to the adjusted BART score because it may be related somewhat more strongly to particular risk-taking behaviors (e.g., Schmitz et al., 2016), which is why we additionally considered this dependent variable in our analyses as a robustness check.

General task representation. We assessed participants' general task representation with the following question: "The question below refers to how the explosion points of the dif-

⁴We ensured a fair payment of at least 8 USD per hour even if participants would have earned less based on their performance. Participants were not previously informed about this policy.

ferent balloons were distributed. Do you believe that the explosion points were clustered around a specific value, or do you believe that the explosion points were randomly distributed across the entire range of the screen?” Participants provided their response using a slider ranging from 0 (labeled “very confident that explosion points were distributed randomly”) to 50 (labeled “very confident that explosion points were clustered”).⁵ In hindsight we realized that the wording of “randomly distributed” might have been ambiguous to some participants, and in study 2 we hence implemented an adapted version of assessing participants’ general task representations.

Beliefs about optimal behavior. To assess participants’ beliefs about the optimal behavior, we asked them to inflate a balloon to the size they expected to yield the maximum payoff in the long run. The instructions read as follows: “Please inflate the balloon to the size that you believe would yield the maximum payoff, were a machine to play this game a thousand times always inflating the balloons to the indicated size.” We prompted participants’ beliefs concerning the optimal behavior only at the end of the task to avoid potential anchoring effects.

Propensity measures. To assess participants’ domain-general risk preferences, we used the GRiPS (Zhang et al., 2018), and the general risk item of the SOEP (e.g., Dohmen et al., 2011). In addition, as risk preferences have been shown to vary across domains (e.g., Weber et al., 2002), we assessed participants’ domain-specific risk-taking propensity using the domain-specific risk items of the SOEP. The exact wording of the items is provided in our preregistration.

Frequency measures. To assess participants’ real-life risk-taking behaviors, we asked them for the frequency with which they had engaged in different activities during the past year. The activities were smoking, drinking, speeding, investing, gambling, and engaging in risky sports (see preregistration for the wording of the items). These activities were chosen to cover domains often assessed in questionnaires of risk-taking propensity (e.g., Blais & Weber, 2006). For each activity, participants could select both the frequency of behavior (from 0 to 100 times) and the desired time frame (per day, per week, per month, or per year).

Statistical analyses

All analyses were conducted using R version 3.6.0 (R Core Team, 2019).

To test Hypothesis 1, we modeled participants’ responses to the question tapping their general task representation (normally vs. uniformly distributed explosion points). To this end, we ran a Bayesian regression model with the group as (non-orthogonal) contrast-coded predictor variable, and the reported beliefs about the distributional form as dependent variables (using the *rstanarm* R package; Goodrich et al., 2018). The contrasts were $\text{BART}_{\text{uniform}}$ vs. the three imple-

mentations of $\text{BART}_{\text{normal}}$, $\text{BART}_{\text{normal-H}}$ vs. $\text{BART}_{\text{normal-M}}$, and $\text{BART}_{\text{normal-M}}$ vs. $\text{BART}_{\text{normal-L}}$.

To test Hypothesis 2, we estimated the differences in means and standard deviations of participants’ beliefs about the optimal behavior in a Bayesian framework. To this end, we used the *BEST* R package (Kruschke, 2013; Kruschke & Meredith, 2018) to fit separate *t*-distributions for the four conditions to participants’ beliefs about the optimal behavior, and then compared the posterior estimates of the means and standard deviations.

To test Hypothesis 3, we estimated the differences in means and standard deviations of participants’ adjusted BART scores in a Bayesian framework. To this end, we again used the *BEST* R package (Kruschke, 2013; Kruschke & Meredith, 2018) to fit separate *t*-distributions for the four conditions to participants’ adjusted BART scores, and then compared the posterior estimates of the means and standard deviations.

To test Hypothesis 4, we report the Pearson correlations of (a) the adjusted BART scores and (b) the total number of explosions per participant with the other measures of risk taking. We computed these correlations separately for the four conditions of the distribution condition in a Bayesian framework using the *BayesFactor* R package (Morey & Rouder, 2018). There were two deviations from our preregistered analysis plan: First, in addition to the adjusted BART score, we used the total number of explosions per participant as a second measure of risk taking, because recent research suggested it to be a potentially better indicator of people’s risk-taking behavior (Schmitz et al., 2016). Second, to make the interpretation of the results regarding Hypothesis 4 more accessible we did not implement the regression models specified in the preregistration but report correlations, which can directly be interpreted as effect sizes. As the frequency ratings indicated some highly skewed distributions, we used binarized versions of these measures in the analyses.

In the analyses, we used the default priors provided by the *rstanarm*, *BEST*, and the *BayesFactor* packages. Specifically, in regression models we used the priors $\mathcal{N}(0, 10)$ for the intercept, and $\mathcal{N}(0, 2.5)$ for the coefficients. In the *t*-tests, we used the priors $\mathcal{N}(\text{mean}(y), \text{sd}(y) * 1000)$ and $\mathcal{U}(\text{sd}(y)/1000, \text{sd}(y) * 1000)$ for μ and σ , and $\mathcal{E}(1/29)$ for ν , with $\nu \geq 1$. Finally, for correlations we used the prior $\text{beta}(3, 3)$.

As suggested by Makowski et al., 2019, we used the ROPE $[-0.1SD_y, 0.1SD_y]$ for testing Hypothesis 1, Hypothesis 2, and Hypothesis 3, and the ROPE $[-0.05, 0.05]$ for testing Hypothesis 4. When reporting parameters, we report the

⁵We preregistered to use a slider ranging from -50 to 50 but accidentally implemented a slider ranging from 0 to 50. Note, however, that only the labels and no numbers were shown to participants. This deviation did thus not affect the appearance of the slider or the interpretations of the results.

median and the 95% HDI of the posterior distribution, as well as the proportion of the posterior distribution that lies within the ROPE (pROPE; note that we interpret the evidence to be conclusive if this value is smaller than .025).

Results

General task representation

In Hypothesis 1, we predicted that at the end of the task participants would believe that the explosion points cluster around a specific value (in line with a normal distribution) rather than that they are uniformly distributed, irrespective of the experimental condition. We intended to interpret ratings larger than the midpoint of the response scale (> 25) as beliefs in line with a normal distribution of explosion points, and ratings below the midpoint of the scale (< 25) as beliefs in line with a uniform distribution of explosion points. As we will discuss below, we realized that this interpretation may not be entirely warranted due to the implemented response format (a more diagnostic response format was thus used in study 2). Yet, according to this definition, only in the BART_{normal-L} did most participants (64.61%) believe that the explosion points were normally distributed, with an average rating of 29.01. In the other implementations, only the minority of 39.69% (BART_{normal-M}), 32.26% (BART_{normal-H}), and 27.72% (BART_{uniform}) of participants had this belief, with average ratings below the midpoint of the scale (i.e., 20.84 in the BART_{normal-M}, 19.15 in the BART_{normal-H}, and 18.06 in the BART_{uniform}). As outlined above, these results have to be interpreted with caution (see discussion section).

Furthermore, as predicted in Hypothesis 1, participants' ratings were increasingly more in line with a normal distribution of explosion points within the BART_{normal}, the smaller the SDs of the explosion points' distributions were: Although there was no conclusive evidence for a difference between participants' ratings in the BART_{normal-M} and the BART_{normal-H} ($b = 1.70$, 95% HDI: [-0.82, 4.22,], pROPE = .382, $d = 0.17$), we found conclusive evidence that participants' ratings in the BART_{normal-L} were higher than in the BART_{normal-M} ($b = 8.16$, 95% HDI: [5.50, 10.61], pROPE $< .001$, $d = 0.63$). Moreover, across the three implementations of the BART_{normal} there was conclusive evidence for higher ratings as compared to in the BART_{uniform} ($b = 4.94$, 95% HDI: [2.87, 7.07], pROPE = .001, $d = 0.37$).

Beliefs about optimal behavior

In Hypothesis 2, we predicted that at the end of the task, participants' beliefs concerning the optimal behavior would consist of a higher number of inflations and less variability between participants in the BART_{normal} as opposed to in the BART_{uniform}. In line with this prediction, participants in the BART_{normal-L} believed the optimal number of inflations to be higher than participants in the BART_{uniform} (see Table 1).

Yet, compared to the BART_{uniform}, there was no conclusive evidence that participants had different beliefs either in the BART_{normal-M} or in the BART_{normal-H}. See Figure 4 for an overview and Table S1 for the estimates of participants' average beliefs.

Furthermore, in line with Hypothesis 2 we found conclusive evidence that the beliefs of participants in the BART_{normal-L} and in the BART_{normal-M} had a smaller variability (i.e., across participants), as compared to the beliefs of participants in the BART_{uniform} (see Table 1). Yet, there was no conclusive evidence whether or not participants in the BART_{normal-H} and in the BART_{uniform} differed concerning the variability of their beliefs.

As the optimal number of inflations varied (i.e., 32, 28, 25, and 25; see Figure 1)⁶ across the four implemented versions of the BART, we also examined the *deviance* between participants' indicated beliefs and the objectively optimal behavior in the respective conditions. When doing so, a similar but even more pronounced pattern in line with Hypothesis 2 emerged. As can be seen in Figure 4, the deviance between participants' beliefs about the optimal behavior and the objectively optimal behavior were consistently larger in the BART_{uniform} than in the three implementations of the BART_{normal} (see Table 1).

Overt risk-taking behavior

In Hypothesis 3, we predicted that participants' adjusted BART scores would be higher and exhibit less variability across participants in the BART_{normal} as opposed to in the BART_{uniform}. Moreover, we predicted that the adjusted BART scores would be higher and exhibit less variability across participants within the BART_{normal}, the lower the standard deviation of the explosion points. In line with this prediction, we found conclusive evidence that, compared to the BART_{uniform}, the adjusted BART scores were higher in all three implementations of the BART_{normal} (see Table 2; see Table S1 for the estimates of participants' adjusted BART scores). Within the BART_{normal} and further in line with Hypothesis 3, there was conclusive evidence that the adjusted BART scores were higher in the BART_{normal-L} than in the BART_{normal-M}. Yet, there was conclusive evidence that the adjusted BART scores in the BART_{normal-M} were lower as compared to those in the BART_{normal-H}.

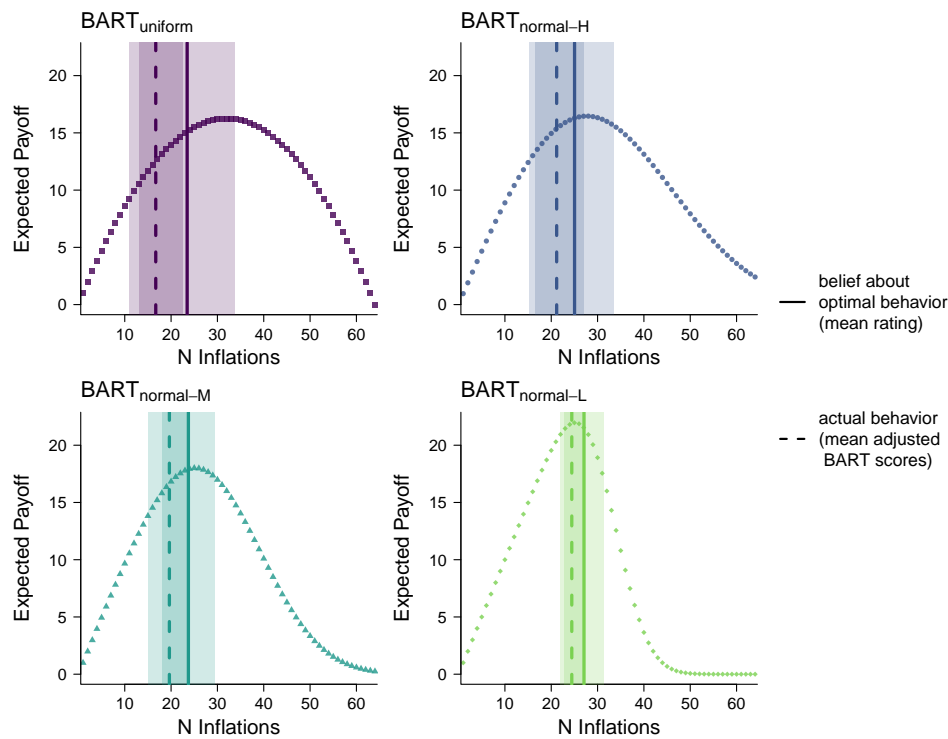
Also in line with Hypothesis 3, there was conclusive evidence that the adjusted BART scores exhibited less variability between participants in the BART_{normal-M} and in the BART_{normal-L}, as compared to in the BART_{uniform} (see Table 2). However, there was no conclusive evidence for

⁶All distributions of explosion points had the same mean of 32, yet lower standard deviations in the BART_{normal} result in a slightly reduced optimal number of inflations. Therefore, the implementations of the BART_{normal} have a lower optimal number of inflations than the BART_{uniform}.

Table 1*Differences in Participants' Beliefs About Optimal Behavior Between Experimental Conditions*

Comparison	Δ [95% HDI]	pROPE	d
<i>Mean beliefs about the optimal behavior</i>			
BART _{normal-H} - BART _{uniform}	1.56 [-0.85, 4.07]	.036	0.16
BART _{normal-M} - BART _{uniform}	0.27 [-1.95, 2.39]	.077	0.11
BART _{normal-L} - BART _{uniform}	3.61 [1.54, 5.64]	< .001	0.38
<i>SD beliefs about the optimal behavior</i>			
BART _{normal-H} - BART _{uniform}	-1.81 [-4.51, 0.80]	.033	-
BART _{normal-M} - BART _{uniform}	-6.03 [-8.30, -3.84]	< .001	-
BART _{normal-L} - BART _{uniform}	-4.69 [-7.05, -2.38]	< .001	-
<i>Deviance of participants' beliefs about optimal behavior from objectively optimal behavior</i>			
BART _{normal-H} - BART _{uniform}	5.55 [3.09, 8.00]	< .001	0.47
BART _{normal-M} - BART _{uniform}	7.27 [5.03, 9.37]	< .001	0.69
BART _{normal-L} - BART _{uniform}	10.60 [8.52, 12.62]	< .001	0.97

Note: The values reported in the first column represent the medians of the posterior distributions and the 95% highest density interval in brackets. The values in the second column (pROPE) represent the proportion of the posterior distribution falling within the region of practical equivalence. The values reported in the third column (d) represent the effect size. Numbers in bold indicate conclusive evidence.

**Figure 4**

Participants' beliefs about the optimal behavior and their actual behavior. Vertical lines indicate the means across participants concerning their beliefs about the optimal behavior (solid lines) and their actual behavior (dashed lines). Shaded areas around the vertical lines indicate the standard deviations of participants' beliefs about the optimal behavior and of their adjusted BART scores.

whether the variability between the BART_{uniform} and the BART_{normal-H} differed. Yet, also in line with Hypothesis 3,

the variability of the adjusted BART scores was lower in the BART_{normal-M} than in the BART_{normal-H}, and lower in the

Table 2*Differences in Overt Risk-Taking Behavior Between Experimental Conditions*

Comparison	Δ [95% HDI]	pROPE	<i>d</i>
<i>Mean adjusted BART scores</i>			
BART _{normal-H} - BART _{uniform}	4.48 [3.25, 5.70]	< .001	0.71
BART _{normal-M} - BART _{uniform}	2.94 [1.84, 4.06]	< .001	0.51
BART _{normal-L} - BART _{uniform}	7.74 [6.73, 8.73]	< .001	1.21
BART _{normal-M} - BART _{normal-H}	-1.54 [-2.64, -0.44]	.002	-0.27
BART _{normal-L} - BART _{normal-M}	4.80 [3.98, 5.63]	< .001	0.77
<i>SD of the adjusted BART scores</i>			
BART _{normal-H} - BART _{uniform}	0.13, [-0.90, 1.17]	.078	-
BART _{normal-M} - BART _{uniform}	-1.31 [-2.37, -0.26]	.004	-
BART _{normal-L} - BART _{uniform}	-3.37 [-4.28, -2.42]	< .001	-
BART _{normal-M} - BART _{normal-H}	-1.44 [-2.50, -0.46]	.001	-
BART _{normal-L} - BART _{normal-M}	-2.06 [-2.95, -1.13]	< .001	-
<i>Deviance of the adjusted BART scores from optimal behavior</i>			
BART _{normal-H} - BART _{uniform}	8.47 [7.22, 9.70]	< .001	1.36
BART _{normal-M} - BART _{uniform}	9.94 [8.82, 11.04]	< .001	1.77
BART _{normal-L} - BART _{uniform}	7.73 [6.72, 8.74]	< .001	2.53
BART _{normal-M} - BART _{normal-H}	1.46 [0.37, 2.57]	.003	0.27
BART _{normal-L} - BART _{normal-M}	4.80 [3.96, 5.62]	< .001	0.77

Note: The values reported in the first column represent the medians of the posterior distributions and the 95% highest density interval in brackets. The values in the second column (pROPE) represent the proportion of the posterior distribution falling within the region of practical equivalence. The values reported in the third column (*d*) represent the effect size. Numbers in bold indicate conclusive evidence.

BART_{normal-L} than in the BART_{normal-M}.

We again also examined the *deviance* between participants' adjusted BART scores and the objectively optimal behavior in the respective conditions. When doing so, a similar but considerably stronger pattern emerged in line with Hypothesis 3: The deviances between the adjusted BART scores and the objectively optimal behavior were much larger in the BART_{uniform} as compared to the three implementations of the BART_{normal} (see Table 2). Moreover, within the BART_{normal} the deviance between the adjusted BART scores and the objectively optimal behavior was larger in the BART_{normal-H} as compared to the in the BART_{normal-M} and larger in the BART_{normal-M} as compared to in the BART_{normal-L}.

Convergent validity

In Hypothesis 4, we predicted that the BART_{normal} would have a higher convergent validity as opposed to the BART_{uniform}. To this end, we tested the correlations of two indicators extracted from the BART (i.e., the adjusted BART score and the total number of explosions per participant) with 14 other measures of risk taking.

Overall, adjusted BART scores were only weakly to moderately related to the other measures (see Figure 5 and Table S2, see Table S11 for descriptive statistics of the different measures), with average correlations of $r = .08$ (BART_{uniform}), $r = -.05$ (BART_{normal-H}), $r = .12$ (BART_{normal-M}), and $r = .04$ (BART_{normal-L}). The total

number of explosions per participant was somewhat more strongly but still weakly related to the other measures, with average correlations of $r = .06$ (BART_{uniform}), $r = -.03$ (BART_{normal-H}), $r = .14$ (BART_{normal-M}), and $r = .05$ (BART_{normal-L}). Moreover, only in the BART_{normal-M} was there a series of measures with conclusive evidence that the correlations were different from 0. Specifically, there was conclusive evidence for associations between the adjusted BART score and GRiPS ($r = .23$), SOEP general ($r = .27$), and SOEP leisure ($r = .24$); and for associations between the total number of explosions and GRiPS ($r = .26$), SOEP general ($r = .30$), SOEP finance ($r = .23$), SOEP health ($r = .20$), and SOEP leisure ($r = .26$). For this reason, we selected the BART_{normal-M} from the three implementations of the BART_{normal} as the focus of our comparison with the BART_{uniform} and report the analyses for the BART_{normal-H} and BART_{normal-L} in the [online Supplemental Material](#) (Section 7.1).

Compared against each other, there were some indications that the BART_{normal-M} exhibited a slightly higher convergent validity with the other measures of risk taking as compared to the BART_{uniform}: The adjusted BART score was more strongly correlated with 11 of the 14 other measures in BART_{normal-M}, and the total number of explosions per participant was more strongly correlated with 12 of the 14 other measures of risk taking (see Figure 5). However, with an average increase of .04 (adjusted BART scores)

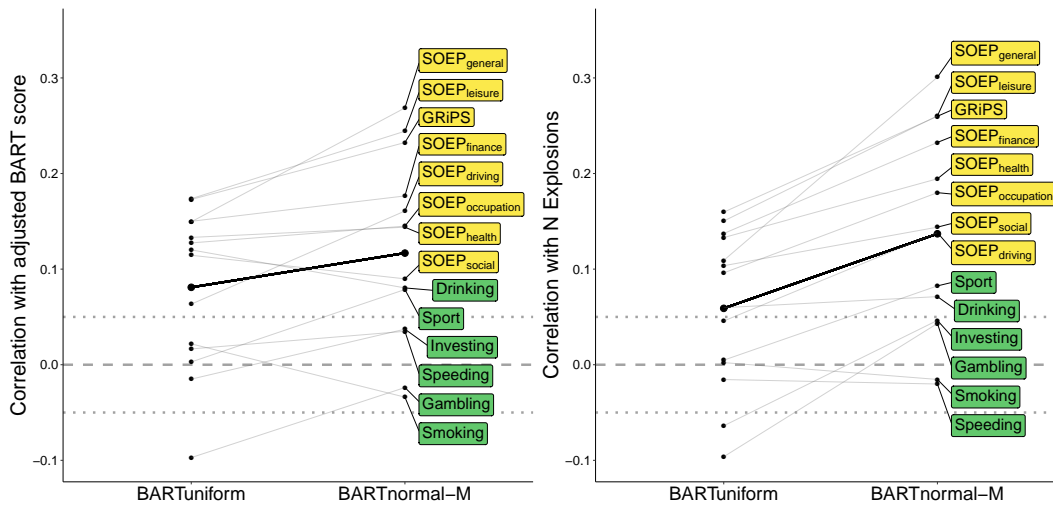


Figure 5

Convergent validity of the BART with other measures of risk taking, separately for the $BART_{uniform}$ and the $BART_{normal-M}$. The left panel shows the correlations based on the adjusted BART scores. The right panel shows the correlations based on the total number of explosions per participant. Propensity measures are depicted in yellow (light gray); frequency measures are depicted in green (dark gray). The dotted lines indicate the boundaries of the region of practical equivalence at $-.05$ and $.05$. The bold black line connects the average correlations of the two task implementations.

and $.08$ (total number of explosions per participant) across the 14 correlations, these differences did not constitute conclusive evidence—although in the most extreme case the correlations almost doubled (i.e., between adjusted BART score and SOEP general) and tripled (i.e., between number of explosions per participant and SOEP general) from the $BART_{uniform}$ to the $BART_{normal-M}$.

Discussion

In study 1 we implemented three new distributions of explosion points to test the potential benefits of employing an improved representative design in the BART. On the one hand, the newly implemented $BART_{normal}$ resulted in several improvements concerning participants' task representations and performance. On the other hand, there was no evident improvement in the task's convergent validity with other measures of risk taking, and all four implemented versions of the BART resulted in similar correlations to those found in earlier studies (e.g., Duckworth & Kern, 2011; Frey et al., 2017; Lauriola et al., 2014; Mishra & Lalumière, 2011). If at all, only the $BART_{normal-M}$ achieved slight improvements in this respect. Yet, the evidence for these increases was not conclusive, and we subsequently tested the convergent validity again in study 2 as a robustness check.

Two specific aspects of these findings warrant further discussion. First, during the assessment of participants' general task representation, we asked whether participants believed

the explosion points in the BART to be *randomly distributed* or to cluster around a specific value. We realized that this assessment might have led to distorted results, for the following two reasons: First, we were not explicit about the meaning of *randomly distributed*. Thus, participants might not necessarily have interpreted this term to mean *uniformly distributed across the whole range*. Second, we prompted participants' beliefs using a continuous slider, with one extreme labeled *randomly distributed* and the other extreme labeled *clustered around one value*. Our intention was to interpret ratings below the midpoint of this scale as evidence that participants' beliefs were in line with a uniform distribution of explosions (and vice versa). Yet, this interpretation is problematic, as any deviation from the left-most rating (i.e., *randomly distributed*) per definition represents some form of clustering—in line with a normal distribution (e.g., a rating of 15 would imply a normal distribution with very wide dispersion). Therefore, we implemented the assessment of participants' general task representation again in study 2, using an improved two-step format as well as making use of visualizations (for details see study 2).

Second, we tested whether the different implementations of the BART resulted in systematically different beliefs about the optimal behavior, as well as in systematically different behaviors (i.e., adjusted BART scores). To this end, we compared these two indicators between the four BART implementations in two ways: by comparing the absolute values, and by comparing the deviance of these values from the ob-

jectively optimal behavior. The latter differed substantially more across the four BART implementations as compared to the former. Yet, although this finding could be interpreted as a strong sign of more accurate learning in the BART_{normal}, we cannot rule out that this pattern also emerged because participants underestimated the average explosion points.

Study 2: Does an Enhanced Representative Design Improve the BART's Test–Retest Reliability?

Study 2 followed our theoretical rationale introduced in study 1, and tested whether a more representative design improves the BART's reliability—in addition to testing the robustness of the findings observed in study 1. Specifically, assuming that people's willingness to take risks remains at least somewhat stable over time (e.g., Frey et al., 2017; Mata et al., 2018), and that people are indeed better able to express their intended degree of risk taking in the BART_{normal} as compared to in the BART_{uniform}, the test–retest reliability of the former should be higher than that of the latter. To test this assumption, we ran a retest of study 1 after about one month. Specifically, we tested the following two hypotheses:

Hypothesis 5: Reliability of beliefs about optimal behavior: Participants' beliefs about the optimal value exhibit a higher test–retest reliability in the BART_{normal} as opposed to in the BART_{uniform}. Moreover, we expected the test–retest reliability within the BART_{normal} to be higher, the lower the standard deviations of the explosion points become.

Hypothesis 6: Reliability of overt risk-taking behavior: There is a higher test–retest reliability of the adjusted BART scores and the total number of explosions per participant in the BART_{normal} as compared to in the BART_{uniform}. Moreover, we expected the test–retest reliability within the BART_{normal} to be higher, the lower the standard deviations of the explosion points become.

Furthermore, we also used study 2 to assess the robustness of the findings observed in study 1, particularly so concerning Hypothesis 1 (i.e., participants' general task representation, where we implemented an improved response format in study 2) and concerning Hypothesis 4 (i.e., the BART's convergent validity with other measures of risk taking and related constructs). Regarding the latter, in study 2 we aimed to test the possibility that the relatively low convergent validity resulted because of our particular selection of additional risk-taking measures, as the BART may also capture related constructs such as impulsivity and sensation seeking (Lauriola et al., 2014; Schmitz et al., 2016; Sharma et al., 2014). To this end, in study 2 we also administered the UPPS scale

(Whiteside & Lynam, 2001; Whiteside et al., 2005), a widely used instrument to tap urgency, lack of premeditation, lack of perseverance, and sensation seeking (for a review and meta-analysis, see Sharma et al., 2014).

Method

Participants and sample characteristics

The 772 participants from study 1 were invited to participate in a retest after an interval of about one month (we sent a maximum of three invitations). We imposed the same inclusion criteria as in study 1. Of the 772 participants from study 1, 671 began with the retest. Of these, 632 met our inclusion criteria and their data were used for the subsequent analyses (46.2% female; $M_{age} = 38.3$, $SD_{age} = 10.9$; highest completed degree: 0.5% no high school, 37.0% high school, 40.7% bachelor, 10.0% master, 10.6% professional, 1.3% doctor; job status: 3.3% student, 11.2% unemployed, 82.6% working, 2.9% retired). On average, study completion took 19 minutes, and on average participants were reimbursed with 4.64 USD. Participants were assigned to the same condition as in study 1 (i.e., of the 632 participants, 157 completed the BART_{uniform}, 158 completed the BART_{normal-H}, 157 completed the BART_{normal-M}, and 160 completed the BART_{normal-L}).

Procedure

The study was again conducted online and participants used their own devices. After providing informed consent, participants completed the BART (i.e., same experimental condition as in study 1; with the same sequence of explosion points). Next, in randomized order, they provided their beliefs about the optimal behavior and reported their general task representation. Then, participants completed, in randomized order, the GRiPS, the assessment of real-life risk-taking behavior, and the SOEP items. At the end of the study, participants completed the UPPS scale and then reported how focused they were during the study, as well as the device they used to complete the study. Finally, participants had the possibility to provide free-text feedback. Screenshots of study 2 are provided at <https://osf.io/kxp8t>.

General task representation

The revised assessment of participants' general task representations was implemented as follows. First, participants received general instructions about the subsequent task and were then presented with two scenarios of distributions of explosion points (i.e., uniform and normal distribution; in randomized order), each of which included an illustration and an explanation of how to read the figures. They then provided a binary rating of whether they believed the explosion points to be uniformly distributed or normally distributed. Finally, participants reported their confidence in their choice

on a slider ranging from 0 (labeled “Not confident at all”) to 50 (labeled “Very confident”). For a detailed formulation of the items, see the preregistration.

Statistical analysis

The retest of Hypothesis 1 and Hypothesis 4 followed the statistical analysis detailed in study 1. To test Hypothesis 1, we first reflected the sign of ratings from participants who had indicated that they believed explosion points to be uniformly distributed and then collapsed the ratings (i.e., resulting in a scale ranging from -50 to 50, with the lower end indicating a high confidence that explosion points were uniformly distributed, and the upper end indicating a high confidence that explosion points were clustered). In the test of Hypothesis 4, we also included the four dimensions of the UPPS scale.

To test Hypothesis 5, we computed the test–retest reliabilities of participants’ beliefs about the optimal behavior, separately for the different BART implementations. We then tested whether there was conclusive evidence that the test–retest reliabilities from the three BART_{normal} implementations were higher than those from the BART_{uniform}. Moreover, we compared the test–retest reliabilities within the BART_{normal} to investigate whether lower variability in the underlying distribution led to higher stability in behavior. Finally, we contrasted the test–retest reliabilities with the coefficient of variation—a standardized measure of dispersion—of the various measures (see [online Supplemental Material Table S10](#)). We conducted the latter analysis to examine possible trade-offs between the measures’ reliability and their potential to capture interindividual differences.

To test Hypothesis 6, we computed the test–retest reliabilities of the adjusted BART scores and the total number of explosions per participant, separately for the different BART implementations. We then tested whether there was conclusive evidence that the test–retest reliabilities of the three BART_{normal} implementations were higher than that of the BART_{uniform}. Moreover, we compared the test–retest reliabilities within the three BART_{normal} implementations to investigate whether lower variability in the underlying distribution leads to higher stability in the behavior. We again contrasted the test–retest reliabilities with the coefficient of variation of the various measures, to analyze possible trade-offs between the measures’ reliability and their potential to capture interindividual differences (see [online Supplemental Material Table S10](#)).

We used the same priors and ROPEs in the analysis of study 2 as we did in study 1.

Results

General task representation

In Hypothesis 1 we predicted that at the end of the task participants would believe that the explosion points cluster around a specific value (in line with a normal distribution) rather than that they are uniformly distributed, irrespective of the experimental condition. As Figure 6 illustrates, this prediction was confirmed: Specifically, 75.3% (BART_{uniform}), 76.3% (BART_{normal-H}), 76.4% (BART_{normal-M}), and 84.2% (BART_{normal-L}) of participants indicated that they believed that the explosion points were clustered, with average confidence ratings of 17.64 in the BART_{uniform}, 19.22 in the BART_{normal-H}, 19.98 in the BART_{normal-M}, and 25.51 in the BART_{normal-L} (on a scale ranging from -50 to 50).

Moreover, as can be seen in Figure 6, there was a trend towards higher confidence in this belief, the narrower the standard deviations of the BART_{normal} became. Yet, there was no conclusive evidence for differences across the tested contrasts between the BART_{uniform} and the BART_{normal} implementations ($b = 3.91$, 95% HDI: [-1.03, 8.70], pROPE = .307, $d = 0.15$), the BART_{normal-M} and the BART_{normal-H} ($b = -0.79$, 95% HDI: [-6.80, 5.35], pROPE = .611, $d = 0.03$), and the BART_{normal-L} and the BART_{normal-M} ($b = -5.49$, 95% HDI: [-11.39, 0.53], pROPE = .174, $d = 0.21$). The pattern that almost no data points are present in the middle of the distribution reveals that most participants were relatively confident in their beliefs about the distributional form of the explosion points.

Convergent validity

In Hypothesis 4 we predicted that the BART_{normal} would have a higher convergent validity with other measures of risk taking than the BART_{uniform}. As the respective evidence was inconclusive in study 1, we tested the convergent validities in study 2 again to investigate whether the observed patterns were robust. To this end, we examined the correlations of two indicators extracted from the BART (i.e., the adjusted BART score and the total number of explosions per participant) with 18 other measures of risk taking.

Overall, participants’ adjusted BART scores were only weakly to moderately related to the other measures (see [Table S6](#)), with average correlations of $r = .08$ (BART_{uniform}), $r = .01$ (BART_{normal-H}), $r = .08$ (BART_{normal-M}), and $r = .05$ (BART_{normal-L}). The total number of explosions per participant exhibited about the same convergent validity as the adjusted BART scores, with average correlations of $r = .08$ (BART_{uniform}), $r = .03$ (BART_{normal-H}), $r = .07$ (BART_{normal-M}), and $r = .06$ (BART_{normal-L}). Moreover, only in the BART_{normal-M} and the BART_{uniform} was there conclusive evidence that some correlations were different from 0. Specifically, in the BART_{uniform} there was conclusive evidence for associations between the adjusted BART scores

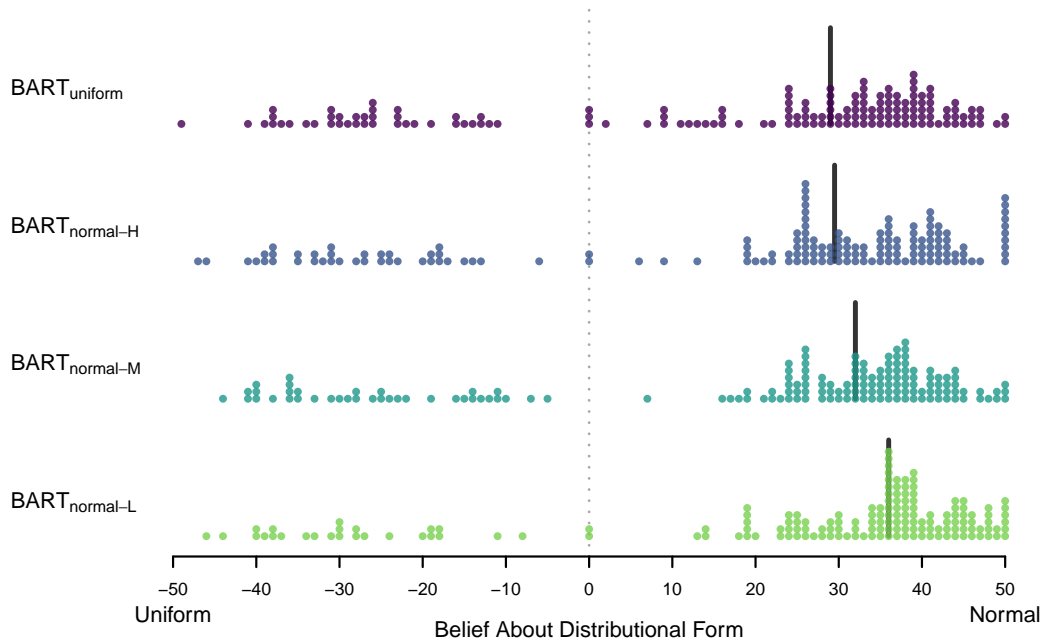


Figure 6

Distributions of participants' beliefs that the balloons' explosion points were uniformly distributed (rating of -50) vs. that they were normally distributed (rating of 50). Beliefs were assessed at the end of the task. Vertical lines indicate the median ratings, separately for the four experimental conditions. The dotted gray line indicates the center of the scale, which corresponds to minimal confidence (i.e., indifference between the two distributional forms).

and the GRiPS ($r = .24$), SOEP general ($r = .22$), and SOEP driving ($r = .22$); and between the total number of explosions per participant and GRiPS ($r = .23$), SOEP general ($r = .24$), and sensation seeking ($r = .22$). In the $BART_{normal-M}$, there was conclusive evidence for associations between the adjusted BART scores and GRiPS ($r = .21$), SOEP leisure ($r = .21$), SOEP social ($r = .26$) and smoking ($r = -.20$); and between the total number of explosions per participant and SOEP general ($r = .20$), and SOEP social ($r = .26$). We again selected the $BART_{normal-M}$ from the three implementations of the $BART_{normal}$ as the focus of our comparison with the $BART_{uniform}$ and report the analyses on $BART_{normal-H}$ and $BART_{normal-L}$ in the [online Supplemental Material](#) (Tables S8 and S9).

Compared to each other, there were no indications that the $BART_{normal-M}$ had a higher convergent validity with the other measures than the $BART_{uniform}$. Specifically, only 8 and 7 of the 18 other measures were more strongly correlated, and 10 and 11 of the 18 other measures were less strongly correlated with the adjusted BART scores and the total number of explosions per participant, respectively. Moreover, the average differences in convergent validity between the $BART_{uniform}$ and the $BART_{normal-M}$ where $\Delta_r = .00$ (adjusted BART scores) and $\Delta_r = -.01$ (total number of explosions per participant) across the 18 correlations.

For the UPPS scale newly included in study 2, the corre-

lations with the adjusted BART scores and the total number of explosions per participant were around the same size as found for the other measures. Specifically, the mean absolute correlations of the four dimensions of the UPPS scale with the adjusted BART scores were $r = .07$ ($BART_{uniform}$), $r = .04$ ($BART_{normal-H}$), $r = .08$ ($BART_{normal-M}$), and $r = .06$ ($BART_{normal-L}$), and those with the total number of explosions per participants were $r = .11$ ($BART_{uniform}$), $r = .05$ ($BART_{normal-H}$), $r = .10$ ($BART_{normal-M}$), and $r = .07$ ($BART_{normal-L}$).

Test-retest reliability of beliefs about optimal behavior

In Hypothesis 5, we predicted that participants' beliefs about the optimal behavior would exhibit a higher test-retest reliability in the $BART_{normal}$ as opposed to in the $BART_{uniform}$, and that within the $BART_{normal}$ implementations, the test-retest reliability would be higher, the lower the standard deviation of the explosion points.

The test-retest reliabilities of participants' beliefs about the optimal behavior were medium to large with $r = .40$ ($BART_{uniform}$), $r = .41$ ($BART_{normal-H}$), $r = .26$ ($BART_{normal-M}$), and $r = .43$ ($BART_{normal-L}$; see also [Figure S5](#) and [Table S10](#)). Contrary to our predictions, there was no conclusive evidence for differences between any of the test-retest reliabilities (see [Table 3](#)).

Table 3

Differences in Test–Retest Reliabilities of BART Indicators Between Experimental Conditions

Implementation	Δ_r [95% HDI]
<i>Belief about the optimal value</i>	
BART _{normal-H} - BART _{uniform}	.00 [-.17, .19]
BART _{normal-M} - BART _{uniform}	-.15 [-.34, .05]
BART _{normal-L} - BART _{uniform}	.03 [-.15, .21]
BART _{normal-M} - BART _{normal-H}	-.15 [-.35, .04]
BART _{normal-L} - BART _{normal-H}	.03 [-.15, .21]
BART _{normal-L} - BART _{normal-M}	.17 [-.01, .37]
<i>Adjusted BART score</i>	
BART _{normal-H} - BART _{uniform}	.14 [.01, .27]
BART _{normal-M} - BART _{uniform}	.07 [-.07, .20]
BART _{normal-L} - BART _{uniform}	-.16 [-.33, .00]
BART _{normal-M} - BART _{normal-H}	-.07 [-.19, .04]
BART _{normal-L} - BART _{normal-H}	-.30 [-.45, -.16]
BART _{normal-L} - BART _{normal-M}	-.23 [-.39, -.08]
<i>Total number of explosions per participant</i>	
BART _{normal-H} - BART _{uniform}	.19 [.04, .33]
BART _{normal-M} - BART _{uniform}	.16 [.01, .31]
BART _{normal-L} - BART _{uniform}	.01 [-.16, .18]
BART _{normal-M} - BART _{normal-H}	-.03 [-.16, .10]
BART _{normal-L} - BART _{normal-H}	-.17 [-.32, -.02]
BART _{normal-L} - BART _{normal-M}	-.14 [-.30, .00]

Note: The reported values represent the medians of the posterior distributions and the 95% highest density interval in brackets. Numbers in bold indicate conclusive evidence.

Test–retest reliability of observed risk-taking behavior

In Hypothesis 6, we predicted that there would be a higher test–retest reliability of the adjusted BART scores and the total number of explosions per participant in the BART_{normal} as compared to in the BART_{uniform}, and that within the BART_{normal} implementations, the test–retest reliability would be higher, the lower the standard deviation of the explosion points.

The test–retest reliabilities of the adjusted BART scores were high in the BART_{uniform} ($r = .59$), the BART_{normal-H} ($r = .73$), and the BART_{normal-M} ($r = .65$), and, surprisingly, somewhat lower in the BART_{normal-L} ($r = .42$). There was conclusive evidence for differences in the test–retest reliabilities between the BART_{normal-H} and the BART_{normal-L} ($\Delta_r = .30$, 95% HDI: [.16, .45]; pROPE < .001), and the BART_{normal-M} and the BART_{normal-L} ($\Delta_r = .23$, 95% HDI: [.08, .39]; pROPE = .007). All other differences represented inconclusive evidence (see Table 3).

Regarding the total number of explosions per participant, we found high test–retest reliabilities in the BART_{normal-H} ($r = .66$) and the BART_{normal-M} ($r = .63$), and somewhat lower ones in the BART_{uniform} ($r = .47$), and the BART_{normal-L} ($r = .48$). There was no conclusive evidence for differ-

ences between the test–retest reliabilities, neither between the BART_{uniform} and the BART_{normal} implementations, nor within the BART_{normal} implementations (see Table 3).

Test–retest reliability of other measures of risk taking

The test–retest reliabilities of the various propensity and frequency measures were similarly high, with average correlations of $r = .68$ and $r = .71$, respectively (see also Figure S5 for an overview of the test–retest reliabilities and the coefficients of variation of all risk-taking measures).

Discussion

In study 2 we tested the robustness of the findings observed in study 1; namely, concerning participants' general task representations and the convergent validity of the BART with other measures of risk taking and related constructs, spanning measures of domain-general and domain-specific risk preference, sensation seeking, impulsivity, and the frequency of specific real-life behaviors. Moreover, we compared the test–retest reliability of the BART_{normal} with that of the BART_{uniform}. As predicted, we observed a strong mismatch between people's general task representation and the stochastic structure of the BART_{uniform}, and this mismatch did not emerge in the BART_{normal}. This corroborates the findings of our reanalyses provided in the first part of this article, namely, that participants' representations of the balloons' explosion points is in line with a normal distribution.

The repeated observation of low convergent validity of the BART as well as its relatively high test–retest reliability call for some discussion; three possibilities have to be considered in this regard. First, low correlations between any two measures may emerge if one of them is unreliable (i.e., the test–retest reliabilities put upper bounds on the correlations between measures; e.g., Kane & Case, 2004). Second, low correlations may emerge if measures fail to capture substantial variation across individuals (i.e., variance restriction). Our results indicated that the BART as well as the other measures performed well in these two respects, with high test–retest reliabilities and high coefficients of variation (i.e., a standardized measure of dispersion; see online Supplemental Material Section 8). Third, low correlations may emerge if measures fail to assess the same underlying constructs or processes involved. In light of the observation that the other risk-taking measures (including measures of impulsivity and sensation seeking) had a high convergent validity between each other (see Figure S4), but not with the BART, our findings imply that the BART may be a relatively reliable task, but it remains unclear *what* it measures (see also our remarks on cognitive modeling in the general discussion).

General Discussion

In this article we investigated the potential benefits of employing the principles of representative design to obtain valid and reliable psychological assessments. We did so by focusing on a widely used behavioral measure of risk taking, the BART. Our primary goal was to test the extent to which adapting an existing task design, by making it more representative, would improve the task's psychometric properties. Such improvements are much needed in various areas of behavioral research (Frey et al., 2017; Lauriola et al., 2014; Lönnqvist et al., 2015; Millroth et al., 2020)—for instance, when investigating the functional neural architecture of risk taking in neuroimaging studies (e.g., Schonberg et al., 2011, 2012; Tisdall et al., 2020). Hence, we reanalyzed data from three previous studies and, based on these findings, adapted the BART's stochastic structure by following the principle of *formal sampling* (Hammond, 1966). Specifically, we changed the distribution of the explosion points from a uniform distribution to a normal distribution—the distribution to be expected from real balloons (Figure 2). Consequently, in two empirical studies we tested whether this adaptation would lead to improvements in participants' beliefs about the task, as well as in the task's psychometric properties. Our main findings can be summarized as follows.

First, our reanalyses of five datasets from three previous studies (Frey et al., 2017; Schürmann et al., 2018; Steiner & Frey, 2020), as well as the results of our experimental studies (in particular study 2; see Figure 6), largely confirmed that the typical implementation of the BART conflicts with participants' beliefs about how explosion points are distributed. Specifically, both before and after having completed the BART, and irrespective of the BART implementation (i.e., BART_{uniform} vs. BART_{normal}), the majority of participants believed that the explosion points clustered around a specific value—in line with a normal distribution, and in line with how real balloons explode (Figure 2).

Second, participants who completed the BART_{normal} (as compared to participants who completed the BART_{uniform}) believed that the optimal behavior was achieved at a higher inflation stage; their beliefs were more closely aligned with the objectively optimal behavior, and also varied less across participants. In terms of their actual behavior, participants' adjusted BART scores were consistently higher, closer to the objectively optimal behavior, and exhibited less variability across participants in the BART_{normal} as compared to in the BART_{uniform}. In short, in the BART_{normal} participants were better able to learn about the optimal behavior, and converged more strongly in doing so—yet without leading to problematic variance restriction—overall suggesting a less noisy learning process. Taken together, these findings confirmed the first three of our hypotheses.

Third and contrary to our expectations, there was no conclusive evidence that these improvements resulted in a sys-

tematic improvement of the BART's convergent validity with other measures of risk taking, nor of its test–retest reliability. Specifically, all four BART implementations correlated only weakly with any of the other risk-taking measures—in line with observations made in previous studies (Duckworth & Kern, 2011; Frey et al., 2017; Lauriola et al., 2014; Mishra & Lalumière, 2011)—whereas the other risk-taking measures (especially the propensity measures) correlated highly with each other. The test–retest reliabilities were relatively high for all implemented versions of the BART as well as for the other risk-taking measures—thus also in line with previous research (Frey et al., 2017; White et al., 2008). This might have left little room for improvement for the BART_{normal} in this respect.

Limitations

All in all, our empirical findings suggest that the BART captures a reliable signal. Yet, our studies indicated that this signal does not consistently tap the constructs of risk preference (in terms of general and domain-specific risk preferences), impulsivity, or sensation seeking, and as such could not reveal *what* this signal reflects. This could be considered a limitation of our study, as yet other psychological constructs (e.g., intelligence; Schmitz et al., 2016) could be assessed in future research, in order to study the role of representative design in fostering the identification of such associations. Relatedly, although several indications suggest that the additional criteria used here to assess risk-taking behaviors are valid (e.g., Dohmen et al., 2011; Eisenberg et al., 2019; Frey et al., 2017; Sharma et al., 2014; Steiner et al., *in Press*), future research may collect further evidence concerning the BART's external validity using yet other measures, and potentially by focusing on extreme groups of specific risk takers (Hopko et al., 2006; Lejuez, Aklin, Jones, et al., 2003; Lejuez et al., 2004).

Moreover, in previous work people's representations of the stochastic structure of the BART have been studied by means of cognitive modeling. This work has put forth important insights and triggered essential discussions on the BART's task design (e.g., concerning whether people may incorrectly adopt a stationary representation of explosion probabilities; Pleskac, 2008; Wallsten et al., 2005, but see Schürmann et al., 2018). In our approach, we did not implement any cognitive modeling analyses but directly prompted participants about their subjective beliefs concerning the distributions of explosion points—following a proof-of-concept recently provided by Schürmann et al. (2018). We followed this route because current models of the BART do not directly account for the underlying task structure at the level we have focused on (i.e., representative design in terms of normal vs. uniform distributions of explosion points), as well as due to a debate concerning parameter recoverability of the state-of-the-art models of the BART (van Ravenzwaaij

et al., 2011). That said, recent developments appear to mitigate the latter issue (Park et al., 2019), and in future work such models (and promising novel variants thereof; Pleskac & Wershba, 2014) may render possible further insights into the cognitive processes involved in the new BART versions presented here.

The Role of Representative Task Design in Psychological Assessment

As introduced in the beginning, representative design refers to “the arrangement of conditions of an experiment so that they represent the behavioral setting to which the results are intended to apply” (Araújo et al., 2007, p. 71). In other words, the experimental stimuli in a task should follow the same stochastic principles (e.g., distributions, intercorrelations) to represent the same or similar cues that are operating in the situations the task is supposed to generalize to (see also, Dhimi et al., 2004). In the ideal case, representative tasks should therefore also tap into the same psychological processes as are present in real-life situations. In the context of risk-taking behaviors, these processes may involve a sensitivity to rewards (e.g., expected benefits, risk conception etc.; Dohmen et al., 2019; Gray, 1982; Kahneman & Tversky, 1979; Weber et al., 2002) and losses (e.g., loss aversion, punishment sensitivity, regret etc.; Gray, 1982; Kahneman & Tversky, 1979; Loomes & Sugden, 1982)—and, depending on the situation, potentially many more factors (e.g., amount of knowledge, affective state, peer influence, competitive pressure; Fischhoff et al., 1978; Frey, 2020; Jellison & Riskind, 1970; Loewenstein et al., 2001; Morrongiello & Lasenby-Lessard, 2007; Phillips et al., 2014).

What does the current observation—that is, that an improved representative design in the BART does not substantially increase its convergent validity with other measures of risk taking—then imply for valid psychological assessments more generally? We see two possibilities in this respect; specifically, representative design may need to be established on two separate levels: First, the behavioral task (here: the BART) needs to be representative of its intended model behavior (here: inflating balloons in real life), requiring adequate abstractions to be used in lab (or online) assessments. Second, the chosen model behavior needs to be representative of the wider class of behavior that is of interest (here: risk-taking behaviors), which relates to the non-trivial issue of selecting an adequate reference class (Hoffrage & Hertwig, 2006).

Concerning representativeness at the *first level*, it may be helpful to draw on two concepts that have been used in research into virtual environments (e.g., flight simulators). The concept of *action fidelity* describes the match between performance in the simulation and performance in the simulated environment (Stoffregen et al., 2003).⁷ Action fidelity implies that stochastic processes and relationships be-

tween variables are similar in the simulated and the real environment—only then will simulated behavior generalize to the respective behavior in reality.⁸ Hence, our adaptation of the BART primarily targeted its action fidelity: Specifically, we employed formal sampling (Dhimi et al., 2004; Hammond, 1966) to close a gap between how the explosions of balloons are distributed in the task and how they are distributed in the real world, making a transfer from task performance to real-life performance more likely in the BART_{normal}. To some extent, this transfer from the abstract virtual environment to the real world may also rest on *experiential fidelity*, which is thought to be present if a person has the feeling of actually being in the simulated environment (Stoffregen et al., 2003). Despite improvements in representative design, even the BART_{normal} might thus have failed to capture relevant psychological processes and respective subjective experiences sufficiently strongly. Although experiential fidelity may not be a *necessary* requirement to achieve action fidelity (Araújo et al., 2007; Moroney et al., 1994; Stoffregen et al., 2003), implementing the BART with loud explosion sounds, or even implementing a BART version with real balloons, may trigger substantially stronger physiological reactions. Yet, it is important to keep in mind the ethical and practical intricacies of such implementations, making their adoption in future assessment contexts unlikely.

Concerning representativeness at the *second level*, a model behavior (e.g., inflating balloons in real life) needs to be representative of the wider class of behaviors that are of interest (e.g., risk-taking behaviors more generally). It has previously been argued that the sequential process of inflating balloons might exhibit properties that are relevant in many risk-taking behaviors, such as the requirement to learn in dynamic environments, the feeling of escalating tension when pursuing additional rewards, and correlated risk-reward structures (Lejuez et al., 2002; Leuker et al., 2018; Pleskac & Hertwig, 2014; Pleskac et al., 2020; Schonberg et al., 2011). The absence of substantial improvements in the BART's external validity (i.e., in response to the elementary stochastic adaptations implemented here) thus hints at another possibility: The model behavior of inflating balloons may simply not represent a wider class of risk-taking behaviors in real life well, thus failing to capture sufficiently many of the psychological processes that are relevant therein. In line with Brunswik's original idea of representative design, we

⁷Task performance can be measured, for example, in terms of transfer effects of training, of completion time needed, or of the variance in performance across trials (e.g., Kozak et al., 1993; Roccio, 1995).

⁸Note that this need not necessarily be the case for a complete real-life behavior from start to end, but can also be the case only for subcomponents of interest. For example, in the case of a flight simulator training, only specific take-off and landing maneuvers may constitute the target behavior, and not necessarily the entire flight.

thus believe that in future work it will be indispensable to first systematize the real-life behaviors of interest—including the involved psychological and structural properties—to then identify promising model behaviors.

A look ahead: Implications for developing new task designs

Our analyses led to two insights for the future development of behavioral tasks. First, under the assumption that the model behaviors of most current tasks (e.g., inflating balloons) do not represent the targeted risk-taking behaviors well, nor capture sufficiently well the relevant psychological processes therein, new model behaviors have to be identified. To this end, ecological analyses will be required to map the actual properties and processes involved in the real-life behaviors of interest, for example, using ecological momentary assessment techniques (e.g., Miller, 2012; Ohly et al., 2010; Trull & Ebner-Priemer, 2013). To illustrate, such momentary assessments could be used to investigate the risks people (have to) take in their lives, what information they consider while doing so, and what the structural properties of the respective environments look like (e.g., Frey, 2020; Pleskac et al., 2020). Based on these insights, respective tasks could be developed with an emphasis on ensuring that the same stochastic structures are present as in the intended model behaviors.

Second, when it comes to the abstraction from identified model behaviors to implementing a behavioral task, it will be important to ensure a sufficiently high level of action fidelity. First and foremost, this implies that the stochastic structure and probabilistic relationships reflect those in the real world. While previous research suggests that very realistic implementations of the model behaviors may not be critical (see Araújo et al., 2007; Moroney et al., 1994; Stoffregen et al., 2003), too abstract tasks might impede action fidelity, such as if they fail to immerse participants in the task (i.e., lack of experiential fidelity). Current behavioral tasks vary widely in this respect, ranging from highly abstract tasks such as multiple price lists (Holt & Laury, 2002) to relatively vivid tasks, such as a driving simulations making use of video clips (Vienna risk-taking test traffic; Hergovich et al., 2007). Further research is needed to examine the extent to which such properties are indeed necessary in order for a task to generalize well to the intended model behavior.

Conclusion

There will be a continued need for behavioral tasks in psychological assessment, including the study of risk-taking behaviors. For instance, in neuroimaging studies behavioral measures are a crucial element to draw valid inferences on the functional neuroanatomy of risk taking. In this article, we reanalyzed five datasets and conducted two experimental studies, aimed at improving the representativeness of the BART. We were arguably successful in doing so with

a simple but important adaptation of one of the BART's most fundamental dimensions: the distribution of explosion points. However, the associated increase in the task's action fidelity—one aspect of representativeness—did not improve its convergent validity, nor its test–retest reliability.

Thus, as long as the model behaviors of current risk-taking tasks do not sufficiently tap the psychological processes that are relevant in real-life risk taking, there is little hope that these tasks can easily be “repaired”, by more closely aligning the task performance with the performance in model behaviors. Therefore, we suggest that future research should aim at developing new behavioral measures by adhering to the principles of representative design at *two levels*: in terms of actual task design, and potentially even more importantly, in terms of an ecologically-guided selection of model behaviors.

Author Contributions

Both authors conceptualized and developed the research. Markus D. Steiner performed the data collection and analysis and drafted the manuscript. Both authors wrote and approved the final version of the article.

References

- Aklin, W., Lejuez, C., Zvolensky, M., Kahler, C., & Gwadz, M. (2005). Evaluation of behavioral measures of risk taking propensity with inner city adolescents. *Behaviour Research and Therapy*, 43(2), 215–228. <https://doi.org/10.1016/j.brat.2003.12.007>
- Araújo, D., Davids, K., & Passos, P. (2007). Ecological validity, representative design, and correspondence between experimental task constraints and behavioral setting: Comment on Rogers, Kadar, and Costall (2005). *Ecological Psychology*, 19(1), 69–78. <https://doi.org/10.1080/10407410709336951>
- Arslan, R. C., Brümmer, M., Dohmen, T., Drewelies, J., Hertwig, R., & Wagner, G. G. (2020). How people know their risk preference. *Scientific Reports*, 10(1), 15365. <https://doi.org/10.1038/s41598-020-72077-5>
- Beauchamp, J., Cesarini, D., & Johannesson, M. (2017). The psychometric and empirical properties of measures of risk preferences. *Journal of Risk and Uncertainty*, 54(3), 203–237. <https://doi.org/10.1007/s11166-017-9261-3>
- Berg, J., Dickhaut, J., & McCabe, K. (2005). Risk preference instability across institutions: A dilemma. *Proceedings of the National Academy of Sciences*, 102(11), 4209–4214. <https://doi.org/10.1073/pnas.0500333102>

- Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2008). How are preferences revealed? *Journal of Public Economics*, 92(8), 1787–1794. <https://doi.org/10.1016/j.jpubeco.2008.04.010>
- Blais, A.-R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1(1), 33–47. <https://doi.org/10.1037/t13084-000>
- Bornovalova, M. A., Daughters, S. B., Hernandez, G. D., Richards, J. B., & Lejuez, C. W. (2005). Differences in impulsivity and risk-taking propensity between primary users of crack cocaine and primary users of heroin in a residential substance-use program. *Experimental and Clinical Psychopharmacology*, 13(4), 311. <https://doi.org/10.1037/1064-1297.13.4.311>
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (Second). University of California Press.
- Campbell, J. A., Samartgis, J. R., & Crowe, S. F. (2013). Impaired decision making on the balloon analogue risk task as a result of long-term alcohol use. *Journal of Clinical and Experimental Neuropsychology*, 35(10), 1071–1081. <https://doi.org/10.1080/13803395.2013.856382>
- Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, 87, 43–51. <https://doi.org/10.1016/j.jebo.2012.12.023>
- Dhimi, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130(6), 959–988. <https://doi.org/10.1037/0033-2909.130.6.959>
- Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3), 522–550. <https://doi.org/10.1111/j.1542-4774.2011.01015.x>
- Dohmen, T., Quercia, S., & Willrodt, J. (2019). Willingness to take risk: The role of risk conception and optimism. *SOEPpapers on Multidisciplinary Panel Data Research*, (1026). <http://hdl.handle.net/10419/195176>
- Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, 45(3), 259–268. <https://doi.org/10.1016/j.jrp.2011.02.004>
- Eisenberg, I. W., Bissett, P. G., Enkavi, A. Z., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, 10(1), 2319. <https://doi.org/10.1038/s41467-019-10301-1>
- Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., & Combs, B. (1978). How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences*, 9(2), 127–152. <https://doi.org/10.1007/BF00143739>
- Frey, R. (2020). Decisions from experience: Competitive search and choice in kind and wicked environments. *Judgment and Decision Making*, 15(2), 282–303. <http://journal.sjdm.org/19/190114/jdm190114.pdf>
- Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10), e1701381. <https://doi.org/10.1126/sciadv.1701381>
- Frey, R., Richter, D., Schupp, J., Hertwig, R., & Mata, R. (2020). Identifying robust correlates of risk preference: A systematic approach using specification curve analysis. *Journal of Personality and Social Psychology*, Advance online publication. <https://doi.org/10.1037/pspp0000287>
- Frey, R., Rieskamp, J., & Hertwig, R. (2015). Sell in may and go away? Learning and risk taking in nonmonotonic decision problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(1), 193–208. <https://doi.org/10.1037/a0038118>
- Gibson, J. J. (1986). *The ecological approach to visual perception*. Lawrence Erlbaum Associates, Inc.
- Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). Rstanarm: Bayesian applied regression modeling via Stan (Version 2.17.4). <http://mc-stan.org/>
- Gray, J. A. (1982). On mapping anxiety. *Behavioral and Brain Sciences*, 5(3), 506–534. <https://doi.org/10.1017/S0140525X00013297>
- Haffke, P., & Hübner, R. (2019). Are choices based on conditional or conjunctive probabilities in a sequential risk-taking task? *Journal of Behavioral Decision Making*. <https://doi.org/10.1002/bdm.2161>
- Hammond, K. R. (1966). Probabilistic functionalism: Egon Brunswik's integration of the history, theory, and method of psychology. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 15–80). Holt, Rinehart and Winston.
- Hanson, K. L., Thayer, R. E., & Tapert, S. F. (2014). Adolescent marijuana users have elevated risk-taking on the balloon analog risk task. *Journal of Psychopharmacology*, 28(11), 1080–1087. <https://doi.org/10.1177/0269881114550352>
- Helfinstein, S. M., Schonberg, T., Congdon, E., Karlsgodt, K. H., Mumford, J. A., Sabb, F. W., Cannon, T. D., London, E. D., Bilder, R. M., & Poldrack, R. A. (2014). Predicting risky choices from brain activity

- patterns. *Proceedings of the National Academy of Sciences*, 111(7), 2470–2475. <https://doi.org/10.1073/pnas.1321728111>
- Hergovich, A., Arendasy, M. E., Sommer, M., & Bognar, B. (2007). The Vienna risk-taking test-traffic: A new measure of road traffic risk-taking. *Journal of Individual Differences*, 28(4), 198–204.
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8), 534–539. <https://doi.org/10.1111/j.0956-7976.2004.00715.x>
- Hoffrage, U., & Hertwig, R. (2006). Which world should be represented in representative design? In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 381–400). Cambridge University Press.
- Holt, C. A., & Laury, S. (2002). Risk aversion and incentive effects. *The American Economic Review*, 92(5), 1644–1655. <https://doi.org/10.1257/000282802762024700>
- Hopko, D. R., Lejuez, C. W., Daughters, S. B., Aklin, W. M., 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(2), 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(4), 416–428. <https://doi.org/10.1177/1073191105278740>
- Jellison, J. M., & Riskind, J. (1970). A social comparison of abilities interpretation of risk-taking behavior. *Journal of Personality and Social Psychology*, 15(4), 375–390. <https://doi.org/10.1037/h0029601>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>
- Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education*, 17(3), 221–240. https://doi.org/10.1207/s15324818ame1703_1
- Knight, F. H. (1921). *Risk, uncertainty, and profit*. Houghton Mifflin Company.
- Kozak, J. J., Hancock, P. A., Arthur, E. J., & Chrysler, S. T. (1993). Transfer of training from virtual reality. *Ergonomics*, 36(7), 777–784. <https://doi.org/10.1080/00140139308967941>
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573–603. <https://doi.org/10.1037/a0029146>
- Kruschke, J. K., & Meredith, M. (2018). BEST: Bayesian estimation supersedes the t-test (Version 0.5.1). <https://CRAN.R-project.org/package=BEST>
- 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(1), 20–36. <https://doi.org/10.1002/bdm.1784>
- Lejuez, C. W., Aklin, W., Daughters, S., Zvolensky, M., Kahler, C., & Gwadz, M. (2007). Reliability and validity of the youth version of the valloon analogue risk task (BART-Y) in the assessment of risk-taking behavior among inner-city adolescents. *Journal of Clinical Child & Adolescent Psychology*, 36(1), 106–111. <https://doi.org/10.1080/15374410709336573>
- 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(1), 26–33. <https://doi.org/10.1037/1064-1297.11.1.26>
- Lejuez, C. W., Simmons, B. L., Aklin, W. M., Daughters, S. B., & Dvir, S. (2004). Risk-taking propensity and risky sexual behavior of individuals in residential substance use treatment. *Addictive Behaviors*, 29(8), 1643–1647. <https://doi.org/10.1016/j.addbeh.2004.02.035>
- Lejuez, C. W., Aklin, W. M., Zvolensky, M. J., & Pedulla, C. M. (2003). Evaluation of the balloon analogue risk task (BART) as a predictor of adolescent real-world risk-taking behaviours. *Journal of Adolescence*, 26(4), 475–479. [https://doi.org/10.1016/S0140-1971\(03\)00036-8](https://doi.org/10.1016/S0140-1971(03)00036-8)
- 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>
- Leuker, C., Pachur, T., Hertwig, R., & Pleskac, T. J. (2018). Exploiting risk–reward structures in decision making under uncertainty. *Cognition*, 175, 186–200. <https://doi.org/10.1016/j.cognition.2018.02.019>
- Li, X., Pan, Y., Fang, Z., Lei, H., Zhang, X., Shi, H., Ma, N., Raine, P., Wetherill, R., Kim, J. J., Wan, Y., & Rao, H. (2019). Test-retest reliability of brain responses to risk-taking during the balloon analogue risk task.

- NeuroImage*, 116495. <https://doi.org/10.1016/j.neuroimage.2019.116495>
- Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127(2), 267–286. <https://doi.org/10.1037/0033-2909.127.2.267>
- Lönnqvist, J.-E., Verkasalo, M., Walkowitz, G., & Wichardt, P. C. (2015). Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. *Journal of Economic Behavior & Organization*, 119, 254–266. <https://doi.org/10.1016/j.jebo.2015.08.003>
- Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, 92(368), 805–824. <https://doi.org/10.2307/2232669>
- Makowski, D., Ben-Shachar, M., & Lüdtke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, 1–8. <https://doi.org/10.21105/joss.01541>
- Mata, R., Frey, R., Richter, D., Schupp, J., & Hertwig, R. (2018). Risk preference: A view from psychology. *The Journal of Economic Perspectives*, 32(2), 155–172. <https://doi.org/10.1257/jep.32.2.155>
- Mata, R., Hau, R., Papassotiropoulos, A., & Hertwig, R. (2012). DAT1 polymorphism is associated with risk taking in the balloon analogue task (BART). *PLoS one*, 7(6), e39135. <https://doi.org/10.1371/journal.pone.0039135>
- Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7(3), 221–237. <https://doi.org/10.1177/1745691612441215>
- Millroth, P., Juslin, P., Winman, A., Nilsson, H., & Lindskog, M. (2020). Preference or ability: Exploring the relations between risk preference, personality, and cognitive abilities. *Journal of Behavioral Decision Making*, 33(4), 477–491. <https://doi.org/10.1002/bdm.2171>
- Mishra, S., & Lalumière, M. L. (2011). Individual differences in risk-propensity: Associations between personality and behavioral measures of risk. *Personality and Individual Differences*, 50(6), 869–873. <https://doi.org/10.1016/j.paid.2010.11.037>
- Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs (Version 0.9.12-4.2). <https://CRAN.R-project.org/package=BayesFactor>
- Moroney, W. F., Hampton, S., Biers, D. W., & Kirton, T. (1994). The use of personal computer-based training devices in teaching instrument flying: A comparative study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 38, 95–99. <https://doi.org/10.1177/154193129403800118>
- Morronegiello, B. A., & Lasenby-Lessard, J. (2007). Psychological determinants of risk taking by children: An integrative model and implications for interventions. *Injury Prevention*, 13(1), 20–25. <https://doi.org/10.1136/ip.2005.011296>
- Mousavi, S., & Gigerenzer, G. (2014). Risk, uncertainty, and heuristics. *Journal of Business Research*, 67(8), 1671–1678. <https://doi.org/10.1016/j.jbusres.2014.02.013>
- Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research. *Journal of Personnel Psychology*, 9, 79–93. <https://doi.org/10.1027/1866-5888/a000009>
- Park, H., Yang, J., Vassileva, J., & Ahn, W.-Y. (2019). The exponential-weight mean-variance model: A novel computational model for the balloon analogue risk task. *PsyArXiv Preprint*. <https://doi.org/10.31234/osf.io/sdzj4>
- Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2014). Rivals in the dark: How competition influences search in decisions under uncertainty. *Cognition*, 133(1), 104–119. <https://doi.org/10.1016/j.cognition.2014.06.006>
- Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 167–185. <https://doi.org/10.1037/0278-7393.34.1.167>
- Pleskac, T. J., Conradt, L., Leuker, C., & Hertwig, R. (2020). The ecology of competition: A theory of risk–reward environments in adaptive decision making. *Psychological Review*, Advance online publication. <https://doi.org/10.1037/rev0000261>
- Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, 143(5), 2000–2019. <https://doi.org/10.1037/xge0000013>
- 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(6), 555–564. <https://doi.org/10.1037/a0014245>
- Pleskac, T. J., & Wershba, A. (2014). Making assessments while taking repeated risks: A pattern of multiple response pathways. *Journal of Experimental Psychology: General*, 143(1), 142–162. <https://doi.org/10.1037/a0031106>
- R Core Team. (2019). R: A language and environment for statistical computing. <http://www.R-project.org>

- Rao, H., Korczykowski, M., Pluta, J., Hoang, A., & Detre, J. A. (2008). Neural correlates of voluntary and involuntary risk taking in the human brain: An fMRI study of the balloon analog risk task (BART). *NeuroImage*, 42(2), 902–910. <https://doi.org/10.1016/j.neuroimage.2008.05.046>
- Roccio, G. E. (1995). Coordination of postural control and vehicular motion: Implications for multimodal perception and simulation of self-motion. In P. Hancock, J. Flach, J. Caird, & K. Vicente (Eds.), *Local applications of the ecological approach to human-machine systems* (pp. 122–181, Vol. 2). Lawrence Erlbaum Associates Publishers.
- Schmitz, F., Manske, K., Preckel, F., & Wilhelm, O. (2016). The multiple faces of risk-taking: Scoring alternatives for the balloon analogue risk task. *European Journal of Psychological Assessment*, 32(1), 17–38. <https://doi.org/10.1027/1015-5759/a000335>
- Schonberg, T., Fox, C. R., Mumford, J. A., Congdon, E., Treppel, C., & Poldrack, R. A. (2012). Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: An fMRI investigation of the balloon analog risk task. *Frontiers in Neuroscience*, 6, 80. <https://doi.org/10.3389/fnins.2012.00080>
- 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(1), 11–19. <https://doi.org/10.1016/j.tics.2010.10.002>
- Schürmann, O., Frey, R., & Pleskac, T. J. (2018). The role of risk perception in dynamic risk-taking behavior. *Journal of Behavioral Decision Making*, 1–12. <https://doi.org/10.1002/bdm.2098>
- Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of “impulsive” behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, 140(2), 374–408. <https://doi.org/10.1037/a0034418>
- Skeel, R. L., Pilarski, C., Pytlak, K., & Neudecker, J. (2008). Personality and performance-based measures in the prediction of alcohol use. *Psychology of Addictive Behaviors*, 22(3), 402–409. <https://doi.org/10.1037/0893-164X.22.3.402>
- Slovic, P. (1962). Convergent validation of risk taking measures. *The Journal of Abnormal and Social Psychology*, 65(1), 68. <https://doi.org/10.1037/h0048048>
- Steiner, M. D., & Frey, R. (2020). Beyond risk preference? Exploring alternative ways to modeling risk taking. *Manuscript in Preparation*.
- Steiner, M. D., Seitz, F. I., & Frey, R. (in Press). Through the window of my mind: Mapping information integration and the cognitive representations underlying self-reported risk preference. *Decision*. <https://doi.org/10.31234/osf.io/sa834>
- Stoffregen, T. A., Bardy, B. G., Smart, L. J., & Pagulayan, R. J. (2003). On the nature and evaluation of fidelity in virtual environments. In L. J. Hettinger & M. W. Haas (Eds.), *Virtual and adaptive environments: Applications, implications, and human performance issues* (pp. 111–128). Lawrence Erlbaum Associates, Inc.
- Tisdall, L., Frey, R., Horn, A., Ostwald, D., Horvath, L., Blankenburg, F., Hertwig, R., & Mata, R. (2020). Brain-behavior associations for risk taking depend on the measures used to capture individual differences. *Frontiers in Behavioral Neuroscience*, 14, 587152. <https://doi.org/10.3389/fnbeh.2020.587152>
- Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, 9, 151–176. <https://doi.org/10.1146/annurev-clinpsy-050212-185510>
- van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*, 55(1), 94–105. <https://doi.org/10.1016/j.jmp.2010.08.010>
- Walasek, L., Wright, R. J., & Rakow, T. (2014). Ownership status and the representation of assets of uncertain value: The balloon endowment risk task (BERT). *Journal of Behavioral Decision Making*, 27(5), 419–432. <https://doi.org/10.1002/bdm.1819>
- Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review*, 112(4), 862–880. <https://doi.org/10.1037/0033-295X.112.4.862>
- Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4), 263–290. <https://doi.org/10.1002/bdm.414>
- White, T. L., Lejuez, C. W., & de Wit, H. (2008). Test-retest characteristics of the balloon analogue risk task (BART). *Experimental and Clinical Psychopharmacology*, 16(6), 565–570. <https://doi.org/10.1037/a0014083>
- 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(4), 669–689. [https://doi.org/10.1016/S0191-8869\(00\)00064-7](https://doi.org/10.1016/S0191-8869(00)00064-7)
- Whiteside, S. P., Lynam, D. R., Miller, J. D., & Reynolds, S. K. (2005). Validation of the UPPS impulsive behaviour scale: A four-factor model of impulsivity.

European Journal of Personality, 19(7), 559–574.
<https://doi.org/10.1002/per.556>

Zhang, D. C., Highhouse, S., & Nye, C. D. (2018). Development and validation of the general risk propensity scale (GRiPS). *Journal of Behavioral Decision Making*, 32, 152–167. <https://doi.org/10.1002/bdm.2102>