Universität Basel

Fakultät für Psychologie

Establishing Construct Validity: The Cases of Risk Preference and Exploratory Factor Analysis

Inauguraldissertation zur Erlangung der Würde eines Doktors der Philosophie vorgelegt der Fakultät für Psychologie der Universität Basel von

Markus Steiner

aus Langnau im Emmental

Basel, 2021

Universität Basel

Fakultät für Psychologie

Genehmigt von der Fakultät für Psychologie auf Antrag von

Prof. Dr. Rui Mata

Prof. Dr. Jörg Rieskamp

Datum des Doktoratsexamen: 23.04.2021

DekanIn der Fakultät für Psychologie

Universität Basel

Fakultät für Psychologie

Erklärung zur wissenschaftlichen Lauterkeit

Ich erkläre hiermit, dass die vorliegende Arbeit ohne die Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel selbstständig verfasst habe. Zu Hilfe genommene Quellen sind als solche gekennzeichnet. Die veröffentlichten oder zur Veröffentlichung in Zeitschriften eingereichten Manuskripte wurden in Zusammenarbeit mit den Koautoren erstellt und von keinem der Beteiligten an anderer Stelle publiziert, zur Publikation eingereicht, oder einer anderen Prüfungsbehörde als Qualifikationsarbeit vorgelegt. Es handelt sich dabei um folgende Manuskripte:

- 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*. Retrieved from: https://psyarxiv.com/sa834/
- Steiner, M.D. & Frey, R. (in press). Representative design in psychological assessment: A case study using the balloon analogue risk task (BART). *Journal of Experimental Psychology: General.* Retrieved from: https://psyarxiv.com/dg4ks/
- Steiner, M. D. & Grieder, S. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53), 2521. doi: 10.21105/joss.02521
- Grieder, S. & Steiner, M. D.[*] (2020). *Algorithmic jingle jungle: A comparison of implementations of principal axis factoring and promax rotation in R and SPSS.* Manuscript submitted for publication*.* Preprint doi:10.31234/osf.io/7hwrm

Basel, 15.02.2021

Markus Steiner

[*] Geteilte Erstautorenschaft.

Acknowledgments

First, I would like to thank Renato Frey for the opportunity to undertake my PhD in his *Ambizione* project, for the many things he taught me during this time, and for the patience with which he met the constant stream of new ideas and thoughts I threw at him and the projects we conducted. Moreover, I thank Rui Mata and Jörg Rieskamp, my PhD supervisors, for their continued support.

I thank Laura Wiles for editing my manuscripts and her patience and flexibility when we were not able to stick to deadlines (sorry...). In addition, I want to thank the whole team of the Center for Cognitive and Decision Sciences for the helpful feedback and for providing an outside perspective on the different projects.

Special thanks go to Dirk Wulff, who has been a mentor of mine for some years now and supported me in many ways, and to Nathaniel Phillips, who has always been there to give me helpful advice when I needed it. I would not be where I am now and certainly would know a lot less, were it not for you guys.

I want to thank my friends and family who have supported me throughout this journey, and, finally and most importantly, thank you to my partner Silvia Grieder for her constant support, for being both a great partner and colleague, and for putting up with my grumbling (I would call it realist [not in a philosophy of science way]) nature.

Contents

Abstract

A crucial precondition for being able to test scientific theories is to clearly define relevant constructs and to validate their assessments. The process of construct validation has been divided into six aspects that focus on different domains of validity evidence, ranging from theoretical considerations to the consequences of assessments and respective score interpretations. In the four manuscripts presented in this dissertation, I focused on several aspects of construct validation in measures of risk preference, as well as on a particular method to investigate the structural aspect of construct validity. Specifically, in manuscript one we investigated the content and substantive aspects of construct validity of self-reported risk preference by focusing on people's cognitive representations of their risk preferences, as well as on potential information integration processes involved during judgment formation. Our results provide further evidence for the validity of assessing risk preference using self reports. In manuscript two, we focused on a different approach to assessing risk preference: behavioral tasks. Specifically, we investigated and aimed to improve the content, substantive, and external aspects of construct validity of the balloon analogue risk task (BART). Adapting the stochastic structure of the BART by following the principles of representative design, we were able to improve the task's content and substantive validity aspects, but not its external validity aspect. Manuscript three presents the *EFAtools* R package that we created to facilitate (a) the process of structural validation of operationalizations, and (b) the comparison of the implementations of a popular exploratory factor analysis (EFA) procedure in R and SPSS. In manuscript four, we then used this package to investigate why this EFA procedure produces differing results when conducted in R than when conducted in SPSS, and whether one of the two implementations should be preferred in construct validation. We found a total of five differences between the two implementations of the EFA procedure that sometimes led to substantial differences in the obtained structural validity evidence. Moreover, we were able to identify an implementation that, on average, maximizes the structural validity evidence obtained with the investigated EFA procedure. With these four manuscripts, this dissertation provides a small, incremental step in the direction of valid assessments of the construct of risk preference, and of improving one of the tools often employed to establish structural validity evidence.

Introduction

As psychologists we strive to understand and describe how the human mind works and how it expresses behavior. For example, we ask how people store information and retrieve it from memory, what kind of dispositions they have and how these shape their interactions and choices, and to what extent such attributes are genetically or environmentally determined, fixed or alterable. To this end, we build and iteratively test theories with the goal to cumulatively advance the science (e.g., Mischel, 2008, 2009). We can think of theories as "specif[ying] interconnections of knowledge" (Gray, 2017, p. 732), or as networks describing relations between observable properties or quantities (manifest variables) and/or psychological constructs (latent variables). Cronbach and Meehl (1955) used the term *nomological networks*, wherein nodes are psychological constructs or manifest variables and edges specify the relations between them. Such nomological networks generate predictions that can be compared to empirical observations. To this end, constructs in the network are operationalized, which is usually done by means of, for example, physiological, self-report, or behavioral measures, that allow us to obtain scores thought to represent people's positions on these constructs. Given these scores, a crucial precondition to being able to test the predictions of a nomological network is for the measures to meet some psychometric properties that serve as indicators that the constructs have been measured well and are indeed usefully (or truthfully) operationalized and conceptualized. In other words, before we can test the interrelations of constructs in the nomological network (edges), there has to be evidence for *construct validity* of the nodes (Cronbach & Meehl, 1955; Messick, 1995). All four manuscripts presented in this dissertation were concerned with this initial step of establishing construct validity. Before I turn to a discussion of what exactly construct validity means, let me expand on ways to think about constructs, as this has important implications for the definition of validity.

In psychological science, constructs are often (at least implicitly) conceptualized in a *reflective* manner (e.g., Borsboom, Mellenbergh, & van Heerden, 2003). A reflective construct is one that is thought to *cause* behavior—that is, an entity that truly exists in the world and thus precedes any measurement of it (Borsboom et al., 2003; Borsboom, Mellenbergh, & van Heerden, 2004). Such entities can consist of processes, sets of processes, or properties of processes (Borsboom et al., 2003; Kovacs & Conway, 2016). A second way of thinking about constructs is in a *formative* manner. A formative construct is nothing more than a summary of a set of observables or other (reflective or formative) constructs. Therefore, it is not causing the manifestation of scores, but is simply an aggregation or a summary thereof—in other words, a purely mathematical entity. Hence, a formative construct does not map onto a real entity. But why is this distinction important for my thesis?

The two interpretations of constructs lead to different definitions of validity—the main focus of this dissertation. If we adopt a realist stance, the typical definition of validity probably most of us would provide if asked—something along the lines of "a test is valid if it measures what it is supposed to measure"—comes closest to the definition applied in such a realist approach (see, Borsboom et al., 2004). Specifically, given a realist interpretation "a test is valid for measuring an attribute if and

only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (Borsboom et al., 2004, p. 1061). Thus, to establish the validity of an assessment, we need to find the function that maps a construct onto some observed score obtained via some operationalization (see also, Kellen, Davis-Stober, Dunn, & Kalish, 2021). In this light, traditional approaches to validation—such as establishing predictive/concurrent validity, construct validity, and content validity (see, Cronbach & Meehl, 1955)—cannot be seen as providing evidence for the validity of a reflective construct, as they are concerned neither with the existence of a construct nor with the relation between the construct and the obtained scores (i.e., points (a) and (b) above; for a detailed discussion, see Borsboom et al., 2004). For the same reason, the popular definition of validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13) is inappropriate when adopting a realist stance. But is following these other approaches to construct validation such as proposed by Messick (1989) a waste of time? I think not. It has to be noted that the field of psychology (and any other field) likely has a long way ahead until we can even come close to something like validating the operationalizations of broad constructs that we currently rely on (i.e., in a realist sense; e.g., Kellen et al., 2021; Meehl, 1978). Until then, we need to employ another approach to evaluate the validity of operationalizations, for example in the manner suggested by Messick (1989). To this end, we could adopt an instrumentalist or constructivist stance, where we can conceptualize constructs in a way that is useful and that can tell us something about the observable world, but that is not necessarily truthful in the realist sense (e.g., for predictive or descriptive purposes; see Yarkoni, 2020; Yarkoni & Westfall, 2017). Even this second, much less ambitious approach brings with it many challenges (e.g., Meehl, 1978; Yarkoni, 2020)—and it is in this framework the manuscripts of my dissertation are positioned. Thus, the definition of validity I will adopt for now is the second one presented above, and is focused on the appropriateness of the interpretations and actions based on test scores (Messick, 1989).

I have briefly mentioned the different subtypes of validity that were initially treated separately (e.g., Cronbach & Meehl, 1955). These have since been integrated into a unified theory of validity (Messick, 1989, 1995), wherein all these subtypes are included in the practice of construct validation. Moreover, it is important to note that "validity is not a property of the test or assessment as such, but rather of the meaning of the test scores [... and that ...] what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails (Cronbach, 1971)" (Messick, 1995, p. 741). In his unified framework of validity, Messick (1989) distinguishes content, substantive, structural, generalizability, external, and consequential aspects of construct validity that serve as validity criteria (see also, Messick, 1995). Given that the manuscripts included in this dissertation focused on four of the six aspects, I will briefly introduce them next.

The content aspect entails the specification of the boundaries and structure of the construct, such as what kind of attributes are expected to be revealed by the

operationalization. Moreover, it is concerned with the *representativeness* of an operationalization for the domain it is supposed to cover (see also, Brunswik, 1955). This first aspect is concerned mostly with theoretical considerations. In contrast, the substantive aspect concerns more empirical considerations regarding the substantive theory and process models of task performance, as well as response consistencies and performance regularities (Loevinger, 1957; Messick, 1995). The structural aspect concerns the consistency between measurement models and what is to be expected based on the involved processes and their dynamic interplay as suggested by, for example, theoretical considerations or task analyses. That is, "the internal structure of the assessment (i.e., interrelations among the scored aspects of task and subtask performance) should be consistent with what is known about the internal structure of the construct domain" (Messick, 1995, p. 746). The generalizability aspect concerns the generalizability of the properties and interpretation of scores to different populations, settings and tasks. External aspects include the relationship between scores from one measure and other measures of the same (in which case relations should be high), or of different constructs (in which case relations should be low), as well as with external criteria (such as indicators for real-life behavior). Thus, the external aspects include what has been called convergent and discriminant validity (Campbell & Fiske, 1959), as well as concurrent/predictive validity (Cronbach & Meehl, 1955). Finally, the consequential aspect includes value implications of scores and consequences of test use (e.g., regarding fairness or bias; Messick, 1989).

The four manuscripts presented here focused on several aspects of construct validation. Two manuscripts focused on operationalizations of the construct of risk preference and examined content, external, and substantive aspects of construct validity. I describe these manuscripts in a first part. In a second part, I then describe the other two manuscripts that were concerned with a procedure often employed to establish the structural aspect of construct validity, namely exploratory factor analysis (EFA).

Part I: Construct Validation of Measures of Risk Preference

We are all faced with numerous decisions every day, most of which involve some degree of risk and uncertainty. In extreme cases, such decisions can determine outcomes like whether we become rich or poor, have longer or shorter lives, or find or lose a partner. Given these profound impacts, it is not surprising that for centuries now the study of risk-taking behaviors has received much attention in psychology and other fields (e.g., Bernoulli, 1738; Kahneman & Tversky, 1979; von Neuman & Morgenstern, 1944, for a historical perspective of the concept of risk, see Aven, 2012; Y. Li, Hills, & Hertwig, 2020). During this time, many theories and models of risk-taking behaviors have been suggested to explain interindividual differences in these behaviors (for an overview, see He, Zhao, & Bhatia, 2020).

The question of why people take risks is often studied through the lens of *risk preference*—that is, people's willingness to take risks—which is thought to be a stable trait (e.g., Frey, Pedroni, Mata, Rieskamp, & Hertwig, 2017; Stigler & Becker, 1977), sometimes with domain-specific components (Frey, Duncan, & Weber, 2020; Frey et al., 2017; Weber, Blais, & Betz, 2002; Wilke et al., 2014). To what extent risk

preference is viewed as a formative or reflective construct is usually not explicitly specified, but the use of terms like *enduring tastes* (e.g., Stigler & Becker, 1977), *appetite for risk* (e.g. Galizzi, Machado, & Miniaci, 2016), *risk attitudes* (e.g., Dohmen et al., 2011) *risk tolerance* (e.g., Linnér et al., 2019), or also *risk preference* itself hints at a mostly reflective interpretation. However, one important issue in this regard is that multiple definitions of *risk*, and therefore also of risk preference, exist (for an overview, see Aven, 2012; Aven, Renn, & Rosa, 2011), and the boundaries between risk preference and related constructs such as impulsivity or sensation seeking are often blurred (Eisenberg et al., 2019; Frey et al., 2017; Sharma, Markon, & Clark, 2014). Clearly, such conceptual clutter can hinder valid operationalizations (in the sense of both Messick, 1989, 1995, and Borsboom et al., 2004), at least regarding certain aspects of construct validity—yet, attempting to solve this issue is beyond what I can hope to achieve in this dissertation. We adopted a conceptualization of risk-taking behaviors often used in psychology; that is, as behaviors that involve potential gains, but also come with the potential for losses (e.g., Mata, Frey, Richter, Schupp, & Hertwig, 2018), and risk preference then is a person's willingness to engage in these kinds of behaviors[1]. To operationalize risk preference, two prominent approaches exist: the *stated preferences* approach, and the *revealed preferences* approach (for a review, see Mata et al., 2018).

In the stated preferences approach, people's risk preferences are assessed using self-report measures. That is, respondents are asked to explicitly state their preferences, usually on some rating scale. These measures have been found to exhibit high test–retest reliabilities and evidence for the external aspect of construct validity (e.g., Dohmen et al., 2011; Frey et al., 2017; Galizzi et al., 2016; Lönnqvist, Verkasalo, Walkowitz, & Wichardt, 2015; Mata et al., 2018). However, concerns have been raised that these self-report measures might show high intercorrelations due to method invariance (i.e., shared variance due to the same method and *response sets* rather than through the same construct being assessed; cf. Cronbach, 1946) and that responses are prone to social desirability biases (e.g., Charness, Gneezy, & Imas, 2013; Harrison & Rutström, 2008; Holt & Laury, 2002). Addressing part of this doubt, specifically, investigating the content and substantive aspects of construct validity of self-reported risk preference, has been the focus of manuscript one.

In the revealed preferences approach, people's risk preferences are inferred based on their choices in behavioral, game-like tasks, such as monetary lotteries or virtual slot machines. These behavioral tasks have sometimes been argued to be the gold standard for assessing risk preference, as they include actual choices that can be incentivized, and are thus thought to be mostly immune to social desirability biases (e.g., Camerer & Hogarth, 1999; Holt & Laury, 2002). Another advantage of this approach is that these very controlled *small worlds* (Savage, 1954) allow for pre-

[1]This definition still does not provide a clear conceptualization of what constitutes risk in these behaviors: the variability in outcomes, the magnitude of a potential loss, the probability of a loss, a combination of these or even additional factors. In the manuscripts I present here, we focused on measures that do not clearly distinguish between these conceptualizations. However, solving this conceptual clutter might still be important in the long run—I will return to this issue in the discussion.

cisely formulated mathematical models that can be subjected to strong tests (Meehl, 1967, 1978)[2]. However, there are a number of issues associated with behavioral tasks, including that they have often been found to exhibit low temporal stability and problems regarding the external aspect of construct validity in terms of low correlations amongst each other and with self-report measures of the same constructs, as well as with measures of relevant real-life risk taking (e.g., Eisenberg et al., 2019; Frey et al., 2017; Lönnqvist et al., 2015; Mata et al., 2018). Addressing especially the content and external aspects of construct validity has been the main focus of manuscript two.

In sum, the two approaches have different proponents and opponents, advantages and disadvantages; yet they are often also combined in a multi-method approach (Dohmen et al., 2011; Frey et al., 2017; Frey, Richter, Schupp, Hertwig, & Mata, 2020; Lejuez et al., 2002; Mishra & Lalumière, 2010). In what follows, I will first describe manuscript one, in which we investigated content and substantive aspects of construct validity of self-reported risk preference. Second, I will describe manuscript two, in which we investigated whether an adaptation of one of the most popular behavioral tasks—the balloon analogue risk task (BART; Lejuez et al., 2002)—might lead to improvements in the content, substantive, and external aspects of the task's construct validity.

Manuscript One: Construct Validation of Self-Reported Risk Preference

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.* Retrieved from `https://psyarxiv.com/sa834/`

As alluded to above, the structural (e.g., Frey et al., 2017), generalizability (e.g., Mata, Josef, & Hertwig, 2016), and external aspects of construct validity of self-reported risk preference (e.g., Dohmen et al., 2011; Galizzi et al., 2016) are relatively well documented. However, comparatively little research has focused on the content and substantive aspects of construct validity, and our goal in this manuscript was to collect evidence in this regard. To this end, we investigated people's cognitive representations underlying these self reports, and strived to describe the possible information integration processes at play. In many scientific studies, in large-scale panel studies as well as in financial institutions, participants' and customers' risk preference is assessed with questions like "Are you generally a person who is willing to take risks or do you try to avoid taking risks?" (this is the general risk item of the German Socio-Economic Panel, SOEP; e.g., Dohmen et al., 2011). We assumed that when coming up with a response to these questions, people retrieve information from memory which they then integrate into a judgment. This sort of internal sampling

[2]Whether the consequences are then actually drawn when a theory gets refuted is a different question. For example, prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) has been refuted many times by strong tests, just like expected utility theory before it (e.g., Birnbaum, 2008; Kellen, Steiner, Davis-Stober, & Pappas, 2020). Yet, it is continued to be widely used. This may reflect the issue that confirmations of theories are often viewed as convincing, but disconfirmations are not (see Beaujean & Benson, 2019).

has been termed *Thurstonian sampling* (e.g., Fiedler & Juslin, 2005; Juslin & Olsson, 1997).

By drawing on research on judgment and decision making that focused mainly on external sampling of information, we hypothesized that three properties of evidence might be especially important when integrating information: First, the *weight of evidence*, which refers to the amount of information pointing in the direction of a specific judgment (Griffin & Tversky, 1992; Kvam & Pleskac, 2016). Second, the *strength of evidence*, which indicates how strongly a piece of information points in the direction of a specific judgment (note that these weights are subjective in the case of internal sampling; Griffin & Tversky, 1992; Kvam & Pleskac, 2016). Third, the *order of evidence*, which refers to the serial position of the individual pieces of evidence (Highhouse & Gallo, 1997; Hogarth & Einhorn, 1992; Yechiam & Busemeyer, 2005). But how can one study the processes by which these kinds of information may be integrated?

We relied on the process-tracing method of *aspect listing* (Johnson, Häubl, & Keinan, 2007; Weber et al., 2007). Therein, people are presented with a judgment or evaluation task, and are asked to concurrently list all the reasons (aspects) that cross their minds during judgment formation. They then indicate their judgment and finally indicate for each aspect whether and how strongly it speaks in favor of a specific judgment (in our case, in favor of seeking risks; we labeled these pro-aspects) or against the respective judgment (contra-aspects). Although aspect listing has mainly been employed with judgments of external objects (Appelt, Hardisty, & Weber, 2011; Johnson et al., 2007; Weber et al., 2007), it has also been used in one of the few studies into the cognitive processes underlying self-report measures (Jarecki & Wilke, 2018; for other studies into self-report measures that relied on similar techniques, see Arslan et al., 2020; Schimmack, Diener, & Oishi, 2002).

The goal of this manuscript was fourfold: First, to investigate whether and how people's self-reported risk preferences can be modeled with a set of cognitive models—that is, to establish substantive validity evidence. To this end, we ran a model comparison of a set of models that incorporated different combinations of the three properties of evidence introduced above (i.e., models of potential information integration processes). Second, to map the content of the listed aspects (i.e., what people thought about) and thus further establish content and substantive validity evidence. Third, to gauge the stability of both the aspects' contents (*aspect stability*), and the aspects' strength of evidence (*evidence stability*). Fourth, to test whether aspect and evidence stability were related to the temporal stability of the self-reported risk preference, which again relates to the substantive aspect of construct validity.

We ran two studies ($N = 250$, and $N = 150$) on Amazon Mechanical Turk (MTurk), the second of which was a within-subjects retest of the first after an interval of about one month. In the two studies, participants were presented with the SOEP general risk item, completed the aspect listing procedure prior to responding to the item, and then indicated the strength of evidence, and provided some information regarding the content of each listed aspect.

We found that people's self-reported risk preferences could indeed be modeled well with a set of cognitive models that take properties of participants' listed aspects as

input. The averaged strength of evidence of these aspects was the best predictor, but the weight of evidence was almost as good. The order of evidence was irrelevant for predictions. These findings apply both to out-of-sample predictions within the same study, as well as across studies. This finding corroborates and extends past research (Jarecki & Wilke, 2018) and provides substantive validity evidence for the employed measure.

Regarding the content of the listed aspects, there was a gap between pro- and contra-aspects: As can be expected, the sentiment of the pro-aspects was much higher (i.e., more positive) as compared to that of contra-aspects. Most participants reported meta-level domain-general statements, often mentioning explicit risk–return tradeoffs (see also, Weber et al., 2002; Weber & Milliman, 1997) and feelings towards taking risks (see also, Bell, 1982; Loewenstein, Weber, Hsee, & Welch, 2001; Loomes & Sugden, 1982; Mellers, Schwartz, Ho, & Ritov, 1997). Moreover, and in line with previous findings, aspects tended to describe active choices and past experiences rather than social comparisons (Arslan et al., 2020; Schimmack et al., 2002; van der Linden, 2014; Weber, 2006). Taken together, these findings speak to the content and substantive validity evidence of self-reported risk preference.

Finally, we found aspect stability to be low (i.e., participants tended to report different aspects across the two time-points), but evidence stability to be high. Moreover, evidence stability was related to the stability of the self-reported risk preference. That is, it might be the case that participants internally sampled from a pool of experiences. Although such a sampling process (especially with small samples as in our study) might lead to different sets of aspects in terms of their contents, if participants tended to have made experiences with similar strengths of evidence, this could explain the evidence stability.

In sum, we found further evidence for the construct validity of self-reports of risk preferences. There is an alignment between the properties of listed aspects and the self-reported risk preferences, the contents of these aspects match what we would expect based on theories of risk taking, and the stability of the properties of the aspects matched the stability of the self-reported risk preference across a one-month period. Although these findings speak to the content and substantive aspects of construct validity of self-report measures of risk preference, it is yet unclear whether we were really able to capture the ongoing information integration processes. In fact, additional data we have meanwhile collected in this regard cast some doubt on our findings. I will return to this issue in the general discussion.

Manuscript Two: Construct Validation of the BART

Steiner, M. D., & Frey, R. (in press). Representative design in psychological assessment: A case study using the balloon analogue risk task (BART). *Journal of Experimental Psychology: General*. Retrieved from `https://psyarxiv.com/dg4ks/`

In contrast to self-report measures of risk preference, behavioral tasks are faced with a different criticism. Although they have been argued to be the gold standard for the assessment of people's risk preference (Beshears, Choi, Laibson, & Madrian,

2008; Charness et al., 2013)—for example, because they are incentive compatible—severe limitations in their test–retest reliabilities as well as the content and external aspects of construct validity have been documented (e.g., Beauchamp, Cesarini, & Johannesson, 2017; Berg, Dickhaut, & McCabe, 2005; Eisenberg et al., 2019; Frey et al., 2017; Lönnqvist et al., 2015; Millroth, Juslin, Winman, Nilsson, & Lindskog, 2020). However, to be able to study risk-taking using, for example, neuroimaging technologies, or in incentive-compatible ways, such behavioral tasks are indispensable (Helfinstein et al., 2014; Rao, Korczykowski, Pluta, Hoang, & Detre, 2008; Schonberg, Fox, & Poldrack, 2011; Tisdall et al., 2020).

In this manuscript, we explored whether one reason for this unsatisfactory state of affairs might be that these tasks are usually designed without following the principles of *representative design* (Brunswik, 1956; Hammond, 1966; for an overview see Araújo, Davids, & Passos, 2007 and Dhami, Hertwig, & Hoffrage, 2004), which is part of the content and substantive aspects of construct validity (Messick, 1995). This concept was introduced by Brunswik and states that experimental stimuli should be sampled or designed such that they represent the environments to which they are supposed to generalize—for example, regarding the stochastic properties of these environments. Based on this concept, we argued that one underlying problem of behavioral tasks might lie in the mismatch between the stochastic structure present in the behavioral tasks and the environments these tasks are supposed to generalize to—and thus, that these problems in the content aspect of construct validity might impede the external aspect of construct validity, as well as temporal stability. To investigate this assumption, we focused on one of the most popular behavioral risk tasks, the BART (Lejuez et al., 2002).

In the BART, participants inflate a number of virtual balloons (usually 30) by repeatedly pressing a button. For every inflation (button press), the virtual balloon increases in size and some fixed amount of money is transferred to a temporary account, that is transferred to a permanent account as soon as the participant decides to stop inflating the current balloon. The balance of this permanent account is paid out at the end of the task. Thus, the goal is to inflate each balloon to as large a size as possible. However, each balloon has an explosion point (i.e., a specific number of inflations) that, when reached, will cause the balloon to explode, in which case the money accrued in the temporary account is lost. This introduces a trade-off, where an optimum number of inflations has to be found in order to maximize the final payoff.

The BART has been argued to exhibit a number of desirable properties also present in real-life behaviors: (a) it is an experience-based task, where properties of the environment have to be learned over time (see also Hertwig, Barron, Weber, & Erev, 2004; Wulff, Mergenthaler-Canseco, & Hertwig, 2018); (b) the risk of a balloon explosion increases with each inflation, leading to a "sense of escalating tension and exhilaration" (Schonberg et al., 2011, p. 16); and (c) risk and reward are positively correlated (e.g., Pleskac, Conradt, Leuker, & Hertwig, 2020; Pleskac & Hertwig, 2014). However, taking a closer look at the typical implementation of the BART, we find that the explosion points are drawn from a uniform distribution—

usually from $\mathcal{U}(1, 128)$[3]. Now, consider if we were to inflate a number of real balloons and inspect the distribution of explosion points: What distributional form would we expect? Arguably we can assume some regularity (as opposed to completely random variability) and some central value around which most balloons would explode—in line with a normal distribution. Indeed, a brief test of this with 100 real balloons showed that the explosion points followed something close to a normal distribution. Hence, the BART's design is not representative in that the stochastic structure of the task environment does not represent the respective real-world environment.

To address this potential shortcoming, we implemented a more representative BART version with a normal distribution of explosion points (the $\text{BART}_{\text{normal}}$, with three versions, all with the same mean but differing standard deviations) and compared it to the task's typical implementation with the uniform distribution of explosion points ($\text{BART}_{\text{uniform}}$, implemented with the same mean as the $\text{BART}_{\text{normal}}$ versions). We hypothesized that the representative task version would improve the accuracy of participants' representations of the task's stochastic structure, of their beliefs about the payoff-maximizing behavior, as well as of their actual behavior in the task. These predictions were based on two assumptions: (a) that the normal distribution is what people expect, should they attempt to make a transfer from their knowledge about real balloons (or given the assumption that many things in the world are normally distributed), and (b) that a normal distribution provides a clearer, less-noisy signal which is easier to learn, due to the more consistent feedback around the mean breaking point. This should facilitate the expression of participants' true preferences due to a better understanding of their current environment (i.e., task structure), and thus improve the association with real-life risk-taking behaviors. Moreover, we assumed that this adaptation would lead to an improvement in the task's temporal stability.

To test these predictions regarding the accuracies of people's representations, beliefs, and behavior, we collected data from 772 participants via MTurk to compare the four BART versions in a between-subjects design (the N per condition ranged between 190 and 197), with a retest after about one month ($N = 632$, ranging between 157 and 160 per condition). Participants first completed one of the four BART versions, then reported whether they believed explosion points to be uniformly or normally distributed, along with a confidence rating, followed by their beliefs about the optimal behavior, and, finally, completed questionnaires assessing both real-life risk-taking behaviors as well as risk propensities in different domains and risk-related constructs.

The results confirmed the first three of our predictions: participants who had completed the $\text{BART}_{\text{normal}}$ exhibited more accurate task representations and beliefs about the optimal behaviors, and displayed more accurate actual behaviors (i.e.,

[3]The initial algorithm to determine explosion points is as follows (see, Lejuez et al., 2002): For each balloon, a vector of $127 \times \mathcal{I}$ (for inflation) and $1 \times \mathcal{E}$ (for explosion) is created. At each inflation, an element from this vector is drawn without replacement. If an $\mathcal{I}$ is drawn, the balloon is inflated, otherwise it explodes. Thus, at each inflation stage where the balloon has not yet exploded on the $i - 1$ preceding trials, the probability that it will explode on the next trial is $p(\mathcal{E}_i) = \frac{1}{C-i+1}$, where C is the maximal capacity of, usually, 128. This leads to a uniform distribution of explosion points.

their scores in the task were closer to the optimal behavior). Strikingly, even in the $\text{BART}_{\text{uniform}}$ condition and after having completed the BART, most participants believed the explosion points to be normally distributed—a clear mismatch between their beliefs and the actual task structure and an invalidation of assumptions of the most popular cognitive models of behavior in the BART (Wallsten, Pleskac, & Lejuez, 2005). However, this improvement in participants' task representations did not translate to improvements in terms of concurrent and convergent validity with other self-report measures of risk preference (both frequency—i.e., real-life risk-taking—and propensity measures) and risk-related constructs, nor to improvements in test–retest reliability. What could be possible reasons for this finding?

We can think about representative design as having to be established on two levels: On the first level, there is the *model behavior*—a behavior that should be representative of the wider class of real-life behaviors we want to generalize to and that is then simulated in a behavioral task. In the case of the BART, the model behavior is the inflation of real balloons in a funfair-like game. Representativeness on this first level is established if the environmental properties—such as stochastic structures in this model behavior—match those present in the real-life behaviors of interest. On the second level, there is the behavioral task—a simulation of the model behavior. This simulation should be representative of the environment in the model behavior and thus should exhibit *action fidelity*—that is, performance in the *simulator* (the task) should match performance in the *simulated* (the model behavior; see Stoffregen, Bardy, Smart, & Pagulayan, 2003). Now, let us assume (a) a task that is representative of its model behavior, and (b) that this model behavior is representative of the wider class of risk-taking behaviors. In this case, the task would also be representative of these real-life behaviors of interest.

We have arguably improved the representativeness of the BART for its model behavior. However, it might be the case that the model behavior of inflating balloons is not representative of the wider class of risk-taking behaviors[4]. But how might we arrive at behavioral tasks that provide valid assessments of people's risk preferences?

If our assumptions are indeed correct and the problem lies at least partly with the model behavior, we first need to identify representative model behaviors. A promising approach to this end is available in the form of ecological momentary assessment techniques (see Miller, 2012; Ohly, Sonnentag, Niessen, & Zapf, 2010; Trull & Ebner-Priemer, 2013), such as the experience sampling method (Hektner, Schmidt, & Csikszentmihalyi, 2007). This would allow us to study the environmental properties and psychological processes involved in the real-life behaviors we ultimately want to predict and understand. From this set of target behaviors, some could then be selected as model behaviors to be simulated in the lab. This way we might reach the goal of arriving at behavioral tasks where we have positive validity evidence in the content, substantive, structural and external aspects of construct validity, and can thus be used to test the theories of decision making, as well as to learn more

[4]In the end, we would have to implement a real-life version of the BART to be sure—an undertaking that would come with great intricacies. For example, inflating 30 real balloons (the typical number of trials in the BART) with a bicycle pump would take about 1 hour. Moreover, a soundproof laboratory (or deaf colleagues) would be needed.

about, for example, the neural underpinnings of risky decisions.

Part II: Establishing Structural Validity Evidence Using Exploratory Factor Analysis

The first part contained two examples of how I investigated (and in one case tried to improve) aspects of construct validity of two specific operationalizations of risk preference, and was thus more content specific. In this second part, I describe two manuscripts that were concerned with a method that can be used to evaluate the structural aspect of construct validity: EFA (introduced by Spearman, 1904).

The goal in establishing structural validity evidence with EFA (and factor analysis in general) is to explore and test the (hypothesized) latent structure of investigated measures, and compare it to what one would expect, for example, based on a task analysis (e.g., C. Li, 2013; Messick, 1995). To this end, most fields of psychology—from clinical psychology (e.g., Derogatis & Cleary, 1977; Dozois, Dobson, & Ahnberg, 1998; Osman, Kopper, Barrios, Osman, & Wade, 1997) to personality psychology (e.g., McCrae & Costa, 1987; Sharma et al., 2014), or the field of intelligence research (e.g., Carroll, 1993; Spearman, 1904)—rely on the framework of factor analysis. For instance, to establish the structure and subscales of the domain-specific risk-taking scale (DOSPERT), Weber et al. (2002), and later Blais and Weber (2006) relied on EFA (see also, Frey, Duncan, & Weber, 2020). Relatedly, to investigate whether risk preference constitutes a uni- or multidimensional construct, Frey et al. (2017) relied on EFA procedures. So, what does factor analysis do? How can we identify the structure (or latent constructs) underlying a set of, for example, questionnaire items as in Weber et al. (2002)?

Factor analysis aims to explain the variance in a larger number of manifest variables with a smaller number of latent factors. EFA constitutes a data-driven approach to factor analysis and can broadly be divided into three substeps[5]: First, the number of latent factors to extract has to be determined. A large number of methods has been proposed to this end (for an overview and comparison, see Auerswald & Moshagen, 2019), which are geared towards different goals (e.g., maximizing verisimilitude or maximizing replicability, see Preacher, Zhang, Kim, & Mels, 2013). This decision is crucial: For example, if we take the publicly available data of Frey, Duncan, and Weber (2020)—which contains responses to the DOSPERT of over 3,000 participants—and subject it to some of the most popular factor-retention criteria, parallel analysis (Horn, 1965) suggests between six and 12 factors (depending on the type), the Kaiser-Guttmann criterion (Guttman, 1954; Kaiser, 1960, 1961) suggests four or seven factors (again, depending on the type), and the scree test (Cattell, 1966) suggests five factors. Now, if we choose to extract, say, a five-factor solution, we would have to test the reliability and validity evidence for each of these five factors separately (e.g., Hubley & Zumbo, 2013). Moreover, the factor structure specifies how the subscales will be assembled. Therefore, this initial decision leads to a great many consequences in scale construction.

[5]Four substeps if we include a prior test of the suitability of a data structure for factor analysis (e.g., Bartlett, 1951).

The second step is to extract the chosen number of factors. A variety of algorithms are available to this end, the most recommended ones being iterative principal axis factoring (PAF) and maximum likelihood estimation (e.g., Costello & Osborne, 2005; Watkins, 2018). These algorithms try to find the set of linear regression equations that can best account for the observed scores based on factor loadings (these are the regression coefficients) and factor scores (people's position on the latent construct—these are the predictors)[6]. In other words, "a common factor model regresses the observed test scores (outcome variables) on the latent factor scores (predictor variables)" (C. Li, 2013, p. 89). The resulting matrix of regression coefficients is called the *loadings* matrix. To predict each of p variables there is one coefficient for each of the m factors, thus the matrix has the dimensions $p \times m$. Often, these loadings are what we are interested in, as they specify the strength of the relation between latent constructs and manifest scores. However, these obtained loadings are frequently hard to interpret.

To facilitate the interpretation of the loadings, in a third step, a factor rotation is performed to seek *simple structure*, where each variable loads saliently[7] onto one, and only one, factor. Two broad types of rotation methods can be distinguished: Orthogonal rotations—where the resulting factors are uncorrelated—and oblique rotations—where the resulting factors are allowed to correlate. It is generally recommended to rely on oblique rotations, as these can account for the complete space of factor intercorrelations (from negative one to one, including zero), whereas orthogonal rotations constrain factor intercorrelations to zero (i.e., orthogonal factors are a special case of oblique factors; e.g., Fabrigar, Wegener, MacCallum, & Strahan, 1999; Gorsuch, 1974; Watkins, 2018). The most popular oblique rotations are promax (Hendrickson & White, 1964) and oblimin (Carroll, 1958; Jennrich & Sampson, 1966, for an overview, see Watkins, 2018). The regression coefficients after oblique rotation are then referred to as *pattern coefficients*.

There are many popular EFA procedures available through the two most popular statistics programs in psychology (Dunn, 2011): R and SPSS. Which of these programs is used should not affect results, and indeed the interchangeable use in publications suggests that no differences between implementations of procedures in the two programs are expected. However, there exists evidence that this interchangeable use is not always justified (e.g., Collins, 2016; del Rio, 2017; GaryStats, 2017; Hodges, Stone, Johnson, Carter, & Lindsey, 2020; krissen, 2018; u/kriesniem, 2018). For instance, in Grieder and Grob (2020), a reviewer asked the authors to verify the EFA results they had obtained with R by rerunning the analysis in SPSS. They followed this suggestion and found the results to differ markedly between the two programs—even though they had specified the same factor extraction and rotation procedure in both programs. Thus, conclusions regarding structural validity evidence drawn from EFA can depend on the software used. The goal of the two manuscripts I describe next was to systematically investigate this issue with the EFA procedure applied in Grieder and Grob (2020)—PAF and promax rotation—and to provide a solution for

[6] Note that the complete right-hand side of this equation is latent; that is, unobserved.

[7] The threshold to determine a loading as salient is usually set at .3 or .4 (e.g., Gorsuch, 1974).

the problem in the form of a freely available, open-source software.

Specifically, the goal of manuscript three was to develop an R package that would allow a fast and systematic test between different EFA procedures, and that would facilitate the process of conducting an EFA by providing convenient meta-level functions. This package (*EFAtools*) was then used in manuscript four, where we systematically compared the R and SPSS implementations of PAF and promax rotation to map (a) how the implementations differ, (b) to what magnitude of differences in results this leads, and (c) whether there exists a best way of implementing these procedures. That is, the goal was to test to what extent we can trust the structural validity evidence provided by these EFA procedures, and how this evidence could be maximized.

Manuscript Three: EFAtools—A Tool for Construct Validation With R

Steiner, M. D. & Grieder, S. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53), 2521. doi: 10.21105/joss.02521

The EFAtools R package implements a set of EFA procedures, including tests of suitability of a data structure for factor analysis, factor-retention criteria, factor extraction and rotation methods, as well as the possibility to compute ω reliability coefficients (e.g., McDonald, 1999). The goal in developing the package was fourfold. First, to provide a collection of easily applicable and modern factor-retention criteria such as comparison data (Ruscio & Roche, 2012), the hull method (Lorenzo-Seva, Timmerman, & Kiers, 2011), or the empirical Kaiser criterion (Braeken & van Assen, 2016), such that the important decision of how many factors to retain could be based on multiple state-of-the-art criteria as suggested by Auerswald and Moshagen (2019). A summary function allows users to run all these criteria with a single function call and shows a summary output that makes the application of this recommendation especially easy. Our second goal was to provide flexible implementations of PAF and promax rotation, such that many different implementations could be run and tested against each other. This allowed us to replicate both the R psych and SPSS implementations, as well as a plethora of further ones. Moreover, we also implemented a large set of other factor extraction and rotation methods. A third goal was to implement a model-averaging function that allows the user to run many implementations at the same time and to obtain an averaged model output that may help gauging the stability of a solution across many implementations. A fourth goal was to provide *C++* implementations of iterative procedures to improve the speed of the analyses, which is especially useful when many EFAs are conducted (e.g., in simulation studies), or when data sets are very large.

With these features, EFAtools facilitates testing the internal structure of measures and can thus be a helpful tool for establishing the structural aspect of construct validity. Moreover, the implementation in a freely available, open-source programming language makes it possible for others to track down details of implementations on a code level, and makes adaptations and further developments much easier as compared to proprietary software. With this package, we then set out to compare the

implementations of PAF and promax rotation between R and SPSS in manuscript four.

Manuscript Four: A Comparison of Implementations of an EFA Procedure in R and SPSS

Grieder, S. & Steiner, M. D. (2020)[8]. *Algorithmic jingle jungle: A comparison of implementations of principal axis factoring and promax rotation in R and SPSS.* Manuscript submitted for publication. Preprint doi: 10.31234/osf.io/7hwrm

Our aim in this manuscript was to identify the reasons why the EFA implementations in R and SPSS produce differing results, as observed in Grieder and Grob (2020)—and the implications for construct validation with these two programs. To this end, we relied on a three-step approach, focusing on PAF and promax as these are among the most popular and robust EFA procedures: First, we compared the implementations on a code/algorithm level, to identify whether the differences were due to programming errors, or valid differences in the implementations. Second, we gauged the magnitude of differences in the results produced by the two implementations across a large collection of real data sets. Third, we ran simulation studies to test whether one implementation outperforms the other, as well as whether there exists an even better implementation that would maximize structural validity evidence obtainable with PAF and promax.

To compare the implementations of PAF and promax in the two programs, we relied on the source code of the implementations in R—more specifically, of the *psych* package (henceforth referred to as the R psych implementation; Revelle, 2020). As SPSS is proprietary software, there was no source code available and we thus relied on the technical manual wherein the algorithms are described (IBM Corp., 2020)[9]. This comparison revealed three differences in the implementations of PAF and two differences in the implementations of the promax procedures. These differences did not constitute programming errors but were either variations in the algorithms that had been suggested in the literature, or just slightly different ways of handling objects and criteria, both of which ways seem valid.

To answer the second point, we factor analyzed 247 data sets from various fields—including the fields of intelligence, personality, and decision making—with both the R psych and SPSS implementations and compared the differences in unrotated loadings and pattern coefficients. This analysis yielded the following main insights: First, the differences after PAF (i.e., between matrices of unrotated loadings) were very small. However, after promax rotation, these differences become larger and, in some cases, substantial. Second, although even after promax rotation the absolute differences in individual loadings often were still relatively small on average, they were large enough to have profound implications in many data sets. Specifically, in 38.4% of these data sets, there was at least one difference in indicator-to-factor correspondences—that is,

[8]Shared first authorship.

[9]Note that we also verified our code by comparing the solutions produced by our code to the solutions produced by the respective implementation in R and SPSS.

differences on which factor the variables loaded saliently (if we only look at data sets with more than one factor extracted, this number was even higher, at 44.4%). In other words, in 38.4% of these data sets the two methods provide diverging evidence for the latent structure, and thus, if we were to develop scales based on these EFAs, the subscales would look different (remember that psychometric properties like reliability and some aspects of validity are then judged per subscale). So, differences exist and are sometimes sizable. But is one implementation preferable over the other?

Analyses of real data sets do not allow comparisons of the two implementations in terms of how accurately these procedures can capture a data-generating process (i.e., the true model), as this process is not known for real data. Yet, ultimately this is the process we try to capture using EFA. Therefore, testing how well the data-generating process can be captured is to establish the validity of such an approach. To overcome this issue, we ran a set of simulation analyses, wherein we created a diverse set of 108 distinct population models (the true data-generating models to recover), from which we then simulated data to subject to EFA. We then not only pitted the two EFA implementations from R psych and SPSS against each other, but included 192 different implementations in a model comparison. These constituted all possible combinations of the differences between the R psych and SPSS implementations, as well as of some additional adaptations suggested in previous literature.

Overall, we found clear and reliable differences in how accurately the implementations were able to recover many of the population models. Regarding only the R psych and SPSS implementations, which implementation was preferable depended on the data structure. Therefore, we cannot make a broad statement of the sort "always use implementation X". A similar picture emerged when considering the complete set of all 192 implementations: We were able to identify an implementation that performed best on average and consists of a mix between the R psych and SPSS implementations. However, it did not consistently perform best across all data structures.

In general we found performance to vary strongly across data structures when comparing the average discrepancies between population models and the factor solutions. This highlights the fact that some data structures are hard to recover in factor analysis, even when we know the true number of factors and when distributional assumptions are fulfilled. Examples are data structures with only few variables per factor, weak pattern coefficients, highly correlated factors, cross-loadings, or variable magnitudes of pattern coefficients (see also de Winter & Dodou, 2012; Gerbing & Hamilton, 1996; Gorsuch, 1974; Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005; MacCallum, Widaman, Zhang, & Hong, 1999; Mulaik, 2010; Tucker & MacCallum, 1997). That is, if such data structures are observed, one should be careful when examining and interpreting the structural validity evidence provided by factor analysis.

Our analyses in this manuscript have shown that the choice of software can impact the structural validity evidence obtained from one of the most popular EFA procedures. Moreover, we were able to identify an implementation that, on average, produces most accurate results, given the data structures considered. However, given that there was no implementation that consistently outperformed all others, a promising alternative approach may also be to employ model averaging to generate an average solution—yet, this remains to be seen in future research.

To summarize, just like the operationalization of a psychological construct, the operationalization of a statistical procedure can influence the validity evidence we obtain. Moreover, just as different scales assessing the same construct are implemented and validated, different algorithmic procedures are implemented and should also be validated.

General Discussion

We can think of theories as nomological networks wherein constructs are the nodes, and the edges are the interrelations between constructs. Testing theories can then be conceptualized as testing the match between the theorized nomological network and the empirically observed one. Yet, only when our proposed constructs are operationalized well and thus exhibit validity evidence to a sufficient degree across the different aspects of construct validity can we tackle the next step of testing theories beyond individual constructs. Otherwise, we risk that different measures of the allegedly same construct may not actually assess the same thing, and thus potentially render tests of theories uninterpretable. Therefore, although the next steps of testing a theory come with substantial intricacies per se (e.g., Kellen et al., 2021; Meehl, 1967, 1978; Yarkoni, 2020), successfully achieving them may be impossible if we do not perform the initial step of construct validation carefully. For example, a test of the risk–return framework (e.g., Weber & Milliman, 1997), where it is assumed that someone's risk preference is determined by their perceived risks and perceived benefits, only makes sense if the involved measures exhibit sufficient evidence for construct validity. How else could we interpret an observed relation between the scores if we are unsure whether the scores represent what we intended to measure?

In the four manuscripts presented in this dissertation, I have focused on studying and improving construct validity in the cases of risk preference and of EFA. The main conclusions from these manuscripts are as follows: (a) Self-reports of risk preference not only exhibit structural, generalizability, and external aspects of construct validity, but also content and substantive aspects. Taken together, these findings likely render these measures useful in an instrumentalist approach. (b) One factor impeding the construct validity evidence of behavioral tasks assessing risk preferences may be the lack of representative design (i.e., a lack in the content and structural aspects of construct validity). Fixing this problem might help us to design behavioral tasks that exhibit sufficient degrees of content, substantive, structural, and external validity evidence. However, additional steps may be necessary for us to be able to create such tasks and until these problems are solved, it might be a sensible approach to refrain from using many behavioral tasks for anything beyond an interest in the tasks themselves. (c) Which statistical software is employed can affect the obtained structural validity evidence and, given current practices of how factor analysis is employed, also how we conceptualize constructs. To maximize the obtained validity evidence from the EFA procedure investigated in manuscript four, the identified best implementation should be applied. Moreover, these differences also highlight that, although it may be perfectly acceptable to rely on these data-driven approaches in an instrumentalist approach in the sense of dimensionality-reduction techniques, it

is important to keep in mind the discussed boundary conditions necessary for these methods to properly function, as well as the kind of evidence they can (and cannot) deliver. Future work is needed along all these lines, both in specific and broader terms.

More specifically, future research on self-reported risk preferences is necessary to test whether the aspects collected in manuscript one really do reflect people's sampling from memory that occurs when they respond to self-report items assessing their risk preference. Although we likely studied their thoughts regarding taking risks, it is unclear whether this information sampling and integration is also the process naturally occurring when no aspect listing precedes the item response (in fact, initial data we have since collected in this direction suggests otherwise). This question could be addressed using think-aloud protocols (Ericsson & Simon, 1980, 1993), that have been shown to not influence task responses (Fox, Ericsson, & Best, 2011). Although our focus was on the construct of risk preference, such an approach would have implications for many other constructs often assessed using self-reports. That is, it could help solving the question of whether people construct judgments of how they see themselves directly when asked, or whether they have stored some value of what kind of person they are in memory (at least for constructs we consider important in everyday lives)[10]. Our findings from manuscript one point to a construction of preferences, however, whether this really is the process taking place has to be studied in future research to corroborate the content and substantive validity evidence of these measures.

Our findings in manuscript two suggested that we may first have to identify representative model behaviors to be able to achieve valid operationalizations of risk preference by means of behavioral tasks. Given that the discussed problems of behavioral tasks also exist in domains other than the study of risk preference (e.g., Duckworth & Kern, 2011; Eisenberg et al., 2019), these findings may also apply to tasks in those other fields. One way to identify representative model behaviors may be to rely on the experience sampling method in combination with think-aloud protocols to study the processes and environmental properties involved in real-life risk-taking behaviors. This would also allow for a more detailed, process-based view on risk-taking behaviors and might thus provide new avenues for theory testing and development on the one hand, and might be beneficial when adopting a prediction focus on the other hand. Moreover, this approach might allow for the creation of tasks that exhibit sufficient degrees of content and substantive validity evidence, which would likely also generalize to external validity evidence.

Finally, regarding the structural validation of scales, model averaging might be a promising alternative to the current use of EFA. Not only different implementations of the PAF and promax procedure as studied here, but also multiple factor-retention and rotation methods could be included in such a procedure to profit from the combination of different properties where the respective methods excel. Moreover, that certain

[10] An alternative explanation that could mimic the retrieval of a stored value could lie in a construction of preferences that does not occur consciously. Such a process could not be investigated using verbal protocol approaches like think-aloud protocols, which might render our situation a tricky one.

data structures could not be recovered well in the EFA simulations highlights the importance of adhering to best practices in scale construction for EFA, and factor analysis in general, to really be able to yield structural validity evidence.

These were some immediate avenues for future research. For the remainder of the discussion I would like to take a broader view. First, regarding the study of risk preference (and most of the risk-related constructs): the apparent conceptual clutter (i.e., lack of a clear, widely accepted and uniformly applied definition of risk) likely impairs our ability to map out clear theories; that is, if we have no clear (functional) definition of the construct, how can the theory around it be precise enough to not fall pray to the scathing criticism offered by Meehl (1978) and to related issues (Kellen et al., 2021; Yarkoni, 2020)? In other words, how can we arrive at a theory from which we can derive precise predictions that help us explain behavior, and that allow for strong tests of the theory? Clearly this problem not only exists with the construct of risk preference and risk-related constructs like impulsivity and sensation seeking, but also in other fields of psychology (see, Meehl, 1978). So, what can we do to address the problems of vague definitions of constructs and theories? I think a clear statement and discussion of our ultimate goals would be a good start, as this determines which road to take. Specifically, if a purely descriptive map of the relations between a set of better or worse defined constructs and/or manifest variables is our goal (cf., Yarkoni, 2020), then the situation is probably not that grave—even though streamlining terminology to the extent possible may still prove helpful. In this *operationalist* approach, a construct would be defined completely by its operationalization in a measure and thus, every measure would make up its own distinct construct (Borsboom et al., 2003)—no more need for construct validation nor explanatory theories in this view. I think little could be gained from such an approach.

In contrast, if the goal were to study and generalize to specific real-life behaviors of interest, it might make sense to define the construct along the lines of these behaviors and then focus on the predictive accuracy of operationalizations (e.g., Yarkoni, 2020; Yarkoni & Westfall, 2017). I think that such an instrumentalist approach could be sensible: It would allow us to clearly define what kind of behaviors we care about, explore what constitute good predictors thereof, and on this basis define constructs, identify their boundaries, processes etc.—in short, embark in construct validation in the sense of Messick (1989, 1995). That is, these constructs would be selected based on their usefulness for the task at hand. It is also this approach for which the findings and methods of the manuscripts included in this dissertation could be usefully applied. Ultimately, this might even lead to sophisticated models akin to those we currently use to model response patterns in behavioral tasks, and therefore could make precise point predictions and even allow for strong tests of theories to be created (e.g., Meehl, 1967)—even though this may not tell us anything about the true state of the world.

Finally, if the goal is to identify the true state of the world, and thus to focus on explanation rather than prediction (even if only at some level of abstraction, given the complexity through the multicausal nature of the world), the task is to disentangle reflective from formative constructs, and come up with process operationalizations based on which the functional mappings onto the measures of the reflective constructs

can be specified. As introduced in the beginning of this dissertation, in this case we need to adopt another definition of validity and can forget about the process of construct validation in the sense of Messick (1989, 1995). Rather, we would have to engage in validation in the sense of Borsboom et al. (2004)—in which case we are up for a long and challenging (and, I think, likely unsuccessful) journey.

Now, *quo vadis*? I can only agree with Yarkoni and Westfall (2017) that an instrumentalist approach focused on prediction is very promising for advancing our field without the need to completely having to turn everything upside down as necessary in a realist approach. Moreover, as I hinted at above, I think adopting a top down approach could prove beneficial, where we start with some real-life behaviors of interest and conduct a task analysis thereof. In a way, that was the start of how the different utility theories were developed—that is, scientists in the 17th and 18th century were interested in optimal solutions to investment and choices in gambles (for a historical overview, see Bernstein, 1996). Yet, I think we may not have performed that step of looking at the real-life behaviors extensively enough afterwards, when we expanded our view beyond monetary gambles. There certainly exist exceptions to this though, and accounting for the ecology in which behaviors take place has been a central focus of certain research programs; thus, I am obviously not the first one to propose this (see, e.g., Brunswik, 1955; Dhami et al., 2004; Goldstein & Gigerenzer, 2002; Leuker, Pachur, Hertwig, & Pleskac, 2018; Pleskac et al., 2020; Simon, 1956). As suggested in the discussion of manuscript two, based on this approach we could start to build measures and study the processes involved therein to eventually arrive at precise predictive models.

These lines of thinking also have implications for the use of factor analysis for the structural validation of constructs. As hinted at above, only in an instrumentalist view does it make sense to embark in construct validation, and thus to apply factor analysis for obtaining structural validity evidence. In this view, factor analysis might provide evidence for the usefulness of applying a specific structure to a set of measures (which to combine and which to separate) and could thus still be beneficial in research—even though some issues remain, such as the practice of drawing individual-level inferences from a between subjects procedure (at least in the way it is usually applied; see Borsboom et al., 2003; Molenaar, 2004; Molenaar & Campbell, 2009).

To conclude, validating operationalizations of hypothesized constructs can consist of different facets, depending on the goal pursued with a given research program. Specifying this goal is crucial, as many decisions concerning the operationalizations, the validation process, the theory tests, and even how we conceptualize theories depend on it. Future work is needed, irrespective of the goal we set and the approach we take, to further advance psychological science.

References

Appelt, K. C., Hardisty, D. J., & Weber, E. U. (2011). Asymmetric discounting of gains and losses: A query theory account. *Journal of Risk and Uncertainty*, *43*(2), 107. doi: 10.1007/s11166-011-9125-1

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. doi: 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. doi: 10.1038/s41598-020-72077-5

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. doi: 10.1037/met0000200

Aven, T. (2012). The risk concept—historical and recent development trends. *Reliability Engineering & System Safety*, *99*, 33–44. doi: 10.1016/j.ress.2011.11.006

Aven, T., Renn, O., & Rosa, E. A. (2011). On the ontological status of the concept of risk. *Safety Science*, *49*(8-9), 1074–1079. doi: 10.1016/j.ssci.2011.04.015

Bartlett, M. S. (1951). The effect of standardization on a $\mathbf{X}^2$ approximation in factor analysis. *Biometrika*, *38*(3/4), 337–344. doi: 10.1093/biomet/38.3-4.337

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. doi: 10.1007/s11166-017-9261-3

Beaujean, A. A., & Benson, N. F. (2019). The one and the many: Enduring legacies of Spearman and Thurstone on intelligence test score interpretation. *Applied Measurement in Education*, *32*(3), 198–215. doi: 10.1080/08957347.2019.1619560

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*(5), 961. doi: 10.1287/opre.30.5.961

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. doi: 10.1073/pnas.0500333102

Bernoulli, D. (1738). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*(1), 23–36. doi: 10.2307/1909829

Bernstein, P. L. (1996). *Against the gods: The remarkable story of risk.* New York: John Wiley & Sons Inc.

Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2008). How are preferences revealed? *Journal of Public Economics*, *92*(8), 1787–1794. doi: 10.1016/j.jpubeco.2008.04.010

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463. doi: 10.1037/0033-295x.115.2.463

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. doi: 10.1037/t13084-000

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. doi: 10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Braeken, J., & van Assen, M. A. L. M. (2016). An empirical Kaiser criterion. *Psychological Methods*, *22*(3), 450. doi: 10.1037/met0000074

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217. doi: 10.1037/h0047470

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (Second ed.). Berkley, CA: University of California Press.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1), 7–42. doi: 10.1023/A:1007850605129

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81. doi: 10.1037/h0046016

Carroll, J. B. (1958). Solution of the oblimin criterion for oblique rotation in factor analysis. *Unpublished manuscript*.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. doi: 10.1207/s15327906mbr0102_10

Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, *87*, 43–51. doi: 10.1016/j.jebo.2012.12.023

Collins, J. (2016). *Multinomial logistic regression in R vs SPSS*. Retrieved from `https://stats.stackexchange.com/questions/189424/multinomial-logistic-regression-in-r-vs-spss`

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, *10*(7), 1–9. doi: 10.7275/jyj1-4868

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. doi: 10.1177/001316444600600405

Cronbach, L. J. (1971). Test validation. In R. Thorndike (Ed.), *Educational measurement* (Second ed., pp. 443–507). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi: 10.1037/h0040957

del Rio, E. (2017). *Comparison of R and SPSS: ANOVA*. Retrieved from `https://medium.com/humansystemsdata/analysis-of-variance-showdown-r-vs-spss-f4e50234a94`

Derogatis, L. R., & Cleary, P. A. (1977). Confirmation of the dimensional structure of the scl-90: A study in construct validation. *Journal of Clinical Psychology*, *33*(4), 981–989. doi: 10.1002/1097-4679(197710)33:4<981::AID-JCLP2270330412>3.0.CO;2-0

de Winter, J. C. F., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710. doi: 10.1080/02664763.2011.610445

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*(6), 959–988. doi: 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. doi: 10.1111/j.1542-4774.2011.01015.x

Dozois, D. J. A., Dobson, K. S., & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory–II. *Psychological Assessment*, *10*(2), 83–89. doi: 10.1037/1040-3590.10.2.83

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. doi: 10.1016/j.jrp.2011.02.004

Dunn, T. J. (2011). The use of 'R' statistical software in psychology research. *PsyPAG Quarterly*, *81*, 10–13. Retrieved from `http://www.psypag.co.uk/`

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. doi: 10.1038/s41467-019-10301-1

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. doi: 10.1037/0033-295x.87.3.215

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272. doi: 10.1037/1082-989x.4.3.272

Fiedler, K., & Juslin, P. (2005). *Information sampling and adaptive cognition*. New York: Cambridge University Press.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. doi: 10.1037/a0021663

Frey, R., Duncan, S., & Weber, E. U. (2020). Towards a typology of risk preference: Four risk profiles describe two thirds of individuals in a large sample of the U.S. population. *PsyArXiv Preprint*. doi: 10.31234/osf.io/yjwr9

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. doi: 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. doi: 10.1037/pspp0000287

Galizzi, M. M., Machado, S. R., & Miniaci, R. (2016). Temporal stability, cross-validity, and external validity of risk preferences measures: Experimental evidence from a UK representative sample. *London School for Economics and Political Science Working Paper*. Retrieved 2016-11-10, from `http://eprints.lse.ac.uk/67554/`

GaryStats. (2017). *Why are SPSS and R producing different results for a cox regression on the same data, with the same model specification?* Retrieved from `https://stats.stackexchange.com/questions/263425/why-are-spss-and-r-producing-different-results-for-a-cox-regression-on-the-same`

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 62–72. doi: 10.1080/10705519609540030

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90. doi: 10.1037//0033-295X.109.1.75

Gorsuch, R. L. (1974). *Factor Analysis*. London: W.B. Saunders Company.

Gray, K. (2017). How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science*, *12*(5), 731–741. doi: 10.1177/1745691617691949

Grieder, S., & Grob, A. (2020). Exploratory factor analyses of the intelligence and development scales–2: Implications for theory and practice. *Assessment*, *27*(8), 1853–1869. doi: 10.1177/1073191119845051

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435. doi: 10.1016/0010-0285(92)90013-R

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161. doi: 10.1007/bf02289162

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). New York: Holt, Rinehart and Winston.

Harrison, G. W., & Rutström, E. E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods. In C. R. Plott & V. L. Smith (Eds.), *Handbook of Experimental Economics Results* (Vol. 1, pp. 752–767). Elsevier. doi: 10.1016/S1574-0722(07)00081-9

He, L., Zhao, W. J., & Bhatia, S. (2020). An ontology of decision models. *Psychological Review*, Advance online publication. doi: 10.1037/rev0000231

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Thousand Oaks, CA: Sage Publications, Inc.

Helfinstein, S. M., Schonberg, T., Congdon, E., Karlsgodt, K. H., Mumford, J. A., Sabb, F. W., ... Poldrack, R. A. (2014). Predicting risky choices from brain activity patterns. *Proceedings of the National Academy of Sciences*, *111*(7), 2470–2475. doi: 10.1073/pnas.1321728111

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*(1), 65–70.

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. doi: 10.1111/j.0956-7976.2004.00715.x

Highhouse, S., & Gallo, A. (1997). Order effects in personnel decision making. *Human Performance*, *10*(1), 31–46. doi: 10.1207/s15327043hup1001_2

Hodges, C. D., Stone, B. M., Johnson, P. K., Carter, J. H., & Lindsey, H. M. (2020). Researcher degrees of freedom and a lack of transparency contribute to unreliable results of nonparametric statistical analyses across SPSS, SAS, Stata, and R. *PsyArXiv Preprint*. doi: 10.31234/osf.io/zem2w

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. doi: 10.1016/0010-0285(92)90002-J

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, *65*(2), 202–226. doi: 10.1177/0013164404267287

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655. doi: 10.1257/000282802762024700

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. doi: 10.1007/bf02289447

Hubley, A. M., & Zumbo, B. D. (2013). Psychmetric characteristics of assessment procedures: An overview. In K. F. Geisingere et al. (Eds.), *APA handbook of testing and assessment in psychology: Test theory and testing and assessment in industrial and organizational psychology* (pp. 3–20). Washington, DC: American Psychological Association.

IBM Corp. (2020). *IBM SPSS Statistics algorithms.* Armonk, NY. Retrieved from `ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/27.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf`

Jarecki, J. B., & Wilke, A. (2018). Into the black box: Tracing information about risks related to 10 evolutionary problems. *Evolutionary Behavioral Sciences*, *12*(3), 230–244. doi: 10.1037/ebs0000123

Jennrich, R. I., & Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika*, *31*(3), 313–323. doi: 10.1007/BF02289465

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 461–474. doi: 10.1037/0278-7393.33.3.461

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*, 344–366. doi: 10.1037/0033-295X.104.2.344

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. doi: 10.2307/1914185

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. doi: 10.1177/001316446002000116

Kaiser, H. F. (1961). A note on Guttman's lower bound for the number of common factors. *British Journal of Statistical Psychology*. doi: 10.1111/j.2044-8317.1961.tb00061.x

Kellen, D., Davis-Stober, C., Dunn, J. C., & Kalish, M. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science*, Advance online publication. Retrieved from `https://psyarxiv.com/3eupv`

Kellen, D., Steiner, M. D., Davis-Stober, C. P., & Pappas, N. R. (2020). Modeling choice paradoxes under risk: From prospect theories to sampling-based accounts. *Cognitive Psychology*, *118*, 101258. doi: 10.1016/j.cogpsych.2019.101258

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*(3), 151–177. doi: 10.1080/1047840x.2016.1153946

krissen. (2018). *Difference between R and SPSS linear model results.* Retrieved from `https://stackoverflow.com/questions/53868465/difference-between-r-and-spss-linear-model-results`

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition*, *152*, 170–180. doi: 10.1016/j.cognition.2016.04.008

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., ... Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The balloon analogue risk task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84. doi: 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. doi: 10 .1016/j.cognition.2018.02.019

Li, C. (2013). Factor analysis of tests and items. In K. F. Geisingere et al. (Eds.), *APA handbook of testing and assessment in psychology: Test theory and testing and assessment in industrial and organizational psychology* (pp. 85–100). Washington, DC: American Psychological Association.

Li, Y., Hills, T., & Hertwig, R. (2020). A brief history of risk. *Cognition*, *203*, 104344. doi: 10.1016/j.cognition.2020.104344

Linnér, R. K., Biroli, P., Kong, E., Meddens, S. F. W., Wedow, R., Fontana, M. A., ... Hammerschlag, A. R. (2019). Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics*, *51*, 245–257. doi: 10.1038/s41588-018-0309-3

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. doi: 10.2466/pr0.1957.3.3.635

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267–286. doi: 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. doi: 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. doi: 10.2307/2232669

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364. doi: 10.1080/00273171.2011.564527

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84. doi: 10.1037/1082-989X.4.1.84

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. doi: 10.1257/jep.32.2.155

Mata, R., Josef, A. K., & Hertwig, R. (2016). Propensity for risk taking across the life span and around the globe. *Psychological Science*, *27*(2), 231–243. doi: 10.1177/ 0956797615617811

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90. doi: 10.1037/0022-3514.52.1.81

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115. doi: 10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. doi: 10.1037/0022-006X.46.4.806

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*(6), 423–429. doi: 10.1111/j.1467-9280.1997.tb00455.x

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (Third ed.). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi: 10.1037/0003-066X.50.9.741

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. doi: 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. doi: 10.1002/bdm.2171

Mischel, W. (2008). The toothbrush problem. *APS Observer*, *21*(11). Retrieved from `https://www.psychologicalscience.org/observer/the-toothbrush-problem`

Mischel, W. (2009). Becoming a cumulative science. *APS Observer*, *22*(1). Retrieved 2021-01-15, from `https://www.psychologicalscience.org/observer/becoming-a-cumulative-science`

Mishra, S., & Lalumière, M. L. (2010). You can't always get what you want: The motivational effect of need on risk-sensitive decision-making. *Journal of Experimental Social Psychology*, *46*(4), 605–611. doi: 10.1016/j.jesp.2009.12.009

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. doi: 10.1207/s15366359mea0204_1

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*(2), 112–117. doi: 10.1111/j.1467-8721.2009.01619.x

Mulaik, S. A. (2010). *Foundations of factor analysis* (Second ed.). Chapman and Hall/CRC.

Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research. *Journal of Personnel Psychology*, *9*, 79–93. doi: 10.1027/1866-5888/a000009

Osman, A., Kopper, B. A., Barrios, F. X., Osman, J. R., & Wade, T. (1997). The Beck Anxiety Inventory: Reexamination of factor structure and psychometric properties. *Journal of Clinical Psychology*, *53*(1), 7–14. doi: 10.1002/(SICI)1097-4679(199701)53:1<7::AID-JCLP2>3.0.CO;2-S

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. doi: 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. doi: 10.1037/xge0000013

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28–56. doi: 10.1080/00273171.2012.710386

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. doi: 10.1016/j.neuroimage.2008.05.046

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research (Version 2.0.12).* Retrieved from `https://CRAN.R-project.org/package=psych`

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*(2), 282–292. doi: 10.1037/a0025697

Savage, L. J. (1954). *The foundations of statistics.* New York, NY: Dover Publications, Inc.

Schimmack, U., Diener, E., & Oishi, S. (2002). Life-satisfaction is a momentary judgment and a stable personality characteristic: The use of chronically accessible and stable sources. *Journal of Personality*, *70*(3), 345–384. doi: 10.1111/1467-6494.05008

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. doi: 10.1016/j.tics.2010.10.002

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. doi: 10.1037/a0034418

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. doi: 10.1037/h0042769

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. doi: 10.2307/1412107

Stigler, G. J., & Becker, G. S. (1977). De gustibus non est disputandum. *The American Economic Review*, *67*(2), 76–90. Retrieved from `https://www.jstor.org/stable/1807222`

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). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Tisdall, L., Frey, R., Horn, A., Ostwald, D., Horvath, L., Blankenburg, F., ... Mata, R. (2020). Brain-behavior associations for risk taking depend on the measures used to capture individual differences. *Frontiers in Behavioral Neuroscience*, *14*, 587152. doi: 10.3389/fnbeh.2020.587152

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151–176. doi: 10.1146/annurev-clinpsy-050212-185510

Tucker, L. R., & MacCallum, R. C. (1997). Exploratory factor analysis. *Unpublished manuscript.*

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. doi: 10.1007/BF00122574

u/kriesniem. (2018). *Different results for Pearson's R as compared to SPSS.* Retrieved from `https://www.reddit.com/r/rstats/comments/a7mzwo/different_results_for_pearsons_r_as_compared_to/`

van der Linden, S. (2014). On the relationship between personal experience, affect and risk perception: The case of climate change. *European Journal of Social Psychology*, *44*(5), 430–440. doi: 10.1002/ejsp.2008

von Neuman, J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

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. doi: 10.1037/0033-295X.112.4.862

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219–246. doi: 10.1177/0095798418771807

Weber, E. U. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change*, *77*(1), 103–120. doi: 10.1007/s10584-006-9060-3

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. doi: 10.1002/bdm.414

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A query-theory account. *Psychological Science*, *18*(6), 516–523.

Weber, E. U., & Milliman, R. A. (1997). Perceived risk attitudes: Relating risk perception to risky choice. *Management Science*, *43*(2), 123–144. doi: 10.1287/mnsc.43.2.123

Wilke, A., Sherman, A., Curdt, B., Mondal, S., Fitzgerald, C., & Kruger, D. J. (2014). An evolutionary domain-specific risk scale. *Evolutionary Behavioral Sciences*, *8*(3), 123–141. doi: 10.1037/ebs0000011

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, *144*(2), 140. doi: 10.1037/bul0000115

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, Advance online publication. doi: 10.1017/S0140525X20001685

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. doi: 10.1177/1745691617693393

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review*, *12*(3), 387–402. doi: 10.3758/bf03193783

Appendix A: Steiner, Seitz, & Frey (in press)

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*. Retrieved from `https://psyarxiv.com/sa834/`

Through the Window of My Mind: Mapping Information Integration and the Cognitive Representations Underlying Self-Reported Risk Preference

Markus D. Steiner[1], Florian I. Seitz[1], Renato Frey[1,2]
[1]University of Basel
[2]Princeton University

Abstract

A person's risk preference may determine significant life outcomes (e.g., in finance or health), and people are therefore routinely asked to report their risk preferences in various scientific and applied contexts. Yet, still little is known concerning the cognitive underpinnings of this judgment-formation process. We ran two studies ($N = 250$, and $N = 150$ in a retest) implementing the process-tracing method of *aspect listing*, to investigate the information-integration processes underlying people's self-reports by means of cognitive modeling (RQ1), as well as to examine people's cognitive representations of their risk preferences (RQ2). Our analyses indicate that interindividual differences in self-reported risk preferences can be modeled well based on the listed aspects' properties of evidence and substantially better than using sociodemographic variables as predictors. Specifically, to render self-reports people appear to integrate the *strength of evidence* of multiple aspects sampled from memory. These aspects further revealed that people's cognitive representation of their risk preferences mostly relate to the magnitudes of outcomes and often to explicit trade-offs between positive and negative outcomes, in line with a risk–return perspective. Crucially, within participants the strength of evidence of the listed aspects remained highly stable across the two studies (RQ3), and changes therein were closely related to changes in self-reported risk preference (RQ4). In sum, our findings provide insight into the cognitive processes of how people render self-reports of their risk preferences, suggest an explanation for the well-documented temporal stability thereof, and thus corroborate the internal validity of this measurement approach.

Keywords: risk preference, self-report measures, process tracing, aspect listing

"Are you generally a person who is willing to take risks or do you try to avoid taking risks?" Chances are that a person is confronted with this or a similar question in numerous settings, such as when discussing private investments with a financial advisor (Balatel et al., 2013; Ferrarini & Wymeersch, 2006) or when taking part in one of the many panel studies that are routinely conducted around the world (e.g., the German Socio-Economic Panel, SOEP; Dohmen et al., 2011; Lejarraga, Frey, Schnitzlein, & Hertwig, 2019). These measurement attempts are not surprising, given that people's risk preferences may shape important life outcomes, such as financial bankruptcy as a consequence of risky investments, or addiction as a consequence of experimenting with substance use. But how do people render judgments concerning their own risk preferences? And can such *stated preferences* indeed be considered valid?

In psychology and the behavioral sciences more generally, self-reports have a long-lasting and successful tradition (Cronbach, 1946; Galton, 1874; Guttman, 1944; Likert, 1932; Thurstone, 1927, 1928). For example, self-report measures were instrumental in the discovery of major constructs such as the Big Five personality dimensions (e.g., McCrae & Costa, 1987) and continue to be an important tool for studying concepts such as grit (e.g., Duckworth, Peterson, Matthews, & Kelly, 2007) or well-being (e.g., Diener, 1984; Kahneman & Deaton, 2010). Crucially, self-report measures not only are easy to implement (e.g., Dohmen et al., 2011; Duckworth & Yeager, 2015) but often also exhibit desirable psychometric properties. To illustrate, in the context of risk preference and closely related constructs, self-report measures were found to have high convergent validity, test–retest reliability, and predictive validity (Beauchamp, Cesarini, & Johannesson, 2017; Duckworth & Yeager, 2015; Frey, Pedroni, Mata, Rieskamp, & Hertwig, 2017; Galizzi, Machado, & Miniaci, 2016; Lönnqvist, Verkasalo, Walkowitz, & Wichardt, 2015; Mata, Frey, Richter, Schupp, & Hertwig, 2018; Rohrer, 2017). By contrast, their behavioral counterparts—that is, game-like tasks such as monetary lotteries, which may be indispensable for applications such as examining the functional neural architecture of risk preference (e.g., Tisdall et al., 2020; Tom, Fox, Trepel, & Poldrack, 2007)—generally tend to be more

---

Markus D. Steiner, Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel; Florian I. Seitz, Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel; Renato Frey, Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel and Behavioral Science for Policy Lab, Andlinger Center for Energy and the Environment, Princeton University.

This work was supported by the Swiss National Science Foundation [grant number PZ00P1_174042 to R.F.]. We thank Olivia Fischer and Samuel Zeiser for their help with the content-ratings of the aspects, Jana Jarecki for helpful comments on an earlier version of this work, and the members of the Center for Cognitive and Decision Sciences for valuable input. We thank Laura Wiles for proofreading. All authors developed the study concepts and contributed to the design. M.D.S. and F.I.S performed the data collection, analysis, and interpretation under the supervision of R.F. M.D.S. and R.F. wrote the manuscript, and F.I.S. provided significant revisions. All authors approved the final version of the article.

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

intricate to implement (see Andreoni & Kuhn, 2019; Pedroni et al., 2017) and often fail to meet fundamental measurement properties (e.g., Beauchamp et al., 2017; Berg, Dickhaut, & McCabe, 2005; Eisenberg et al., 2019; Frey et al., 2017; Lönnqvist et al., 2015; Mata et al., 2018; Steiner & Frey, in press).

Given the widespread adoption of self-report measures of risk preference in the behavioral sciences, it is surprising that there has been hardly any effort to systematically examine the cognitive processes and representations underlying people's self-reports (for exceptions, see Arslan et al., 2020; Jarecki & Wilke, 2018) and to thus shed some light onto the potential origins of these measures' desirable psychometric properties. Hence, several important questions remain largely unaddressed: What kind of evidence do people rely on during their judgment-formation process? What are the qualitative and quantitative properties of this process? And do cognitive explanations exist for the observation that people's self-reported risk preferences remain highly stable across time (e.g., Frey et al., 2017; Lönnqvist et al., 2015; Mata et al., 2018)? The goal of this article is to address these questions and "unpack" people's self-reports of their risk preferences, by modeling the information-integration processes underlying such self-reports, and thus testing the internal validity of this measurement approach.

The Psychology of Judgment Formation

Several streams in cognitive psychology assume judgment formation to rest on some form of *internal* or *external* information-sampling process. External information sampling (i.e., *Brunswikian sampling*; Fiedler & Juslin, 2005; Juslin & Olsson, 1997) has been extensively studied, often with the finding that observed samples predict people's choices and behaviors well (e.g., Fiedler, Renn, & Kareev, 2010; Hertwig, Barron, Weber, & Erev, 2006; Lindskog, Winman, & Juslin, 2013). In these investigations, computational models constitute a tool to systematically study the links between the external samples that participants observed and their choices or judgments—by formalizing the cognitive processes involved in information use and integration (e.g., Frey, Mata, & Hertwig, 2015; Frey, Rieskamp, & Hertwig, 2015; Kellen, Pachur, & Hertwig, 2016; Yechiam & Busemeyer, 2005).

When facing the task of providing a self-report, one typically cannot rely on external information but instead has to draw internal samples of one's own past behaviors and experiences (i.e., *Thurstonian sampling*; Bem, 1967; Fiedler & Juslin, 2005; E. J. Johnson, Häubl, & Keinan, 2007; Juslin & Olsson, 1997). This process may involve three broad stages, each involving different cognitive processes: First, information has to be retrieved from memory. Although memory retrieval has been a central assumption in models of *survey cognition* (e.g., Duckworth & Yeager, 2015; Jobe, 2000, 2003; Schwarz & Oyserman, 2001), concrete properties of this process have rarely been specified (for an overview, see Jobe & Herrmann, 1996; Koriat, Goldsmith, & Pansky, 2000; Tourangeau, Rips, & Rasinski, 2000). Second, the information retrieved from memory has to be integrated into an internal representation. Third and finally, the result of this information-integration process has to be rendered into a concrete output, for instance, mapping onto a specific response format (e.g., a

Likert scale).

The focus of this article lies on the second stage; that is, the information-integration processes underlying people's self-reports. The respective cognitive processes have only rarely been studied systematically (for an exception, see Jarecki & Wilke, 2018) but information-integration processes are naturally paramount in research on judgment and decision making more generally (e.g., Dawes & Corrigan, 1974; Gigerenzer & Goldstein, 1996; Hastie & Park, 1986; Payne, Bettman, & Johnson, 1988). This line of research has identified three basic properties of evidence, which may also be of importance when people integrate information to render a self-report of their risk preferences: First, in their work on confidence judgments, Griffin and Tversky (1992) referred to the *weight of evidence* as the amount of information taken into account during a particular judgment (see also Kvam & Pleskac, 2016). Specifically, in their study design in the context of fairness assessments of biased coins (i.e., external samples), the weight of evidence referred to how many times a coin was spun and hence, the number of outcomes observed (i.e., sample size). Translated into the process of self-reporting one's risk preference, the weight of evidence may consist of how many pieces of information are retrieved from memory and either speak pro or contra risk taking. Indeed, initial evidence suggests that the weight of evidence of retrieved information may play an important role in the context of rendering self-reports (see introduction to study 1).

Second, the *strength of evidence* refers to the extremeness of the available information; that is, how strongly a particular piece of information supports a certain judgment (Griffin & Tversky, 1992; Kvam & Pleskac, 2016; see also Koriat, 1993). In the work of Griffin and Tversky (1992), the strength of evidence was defined as how strongly the bias showed up across the entire sample of spun coins (i.e., effect size). Whereas in this example a single coin spin always yields equally strong evidence (i.e., in one or the other direction), in other contexts single pieces of information vary in terms of their strength of evidence (e.g., Hertwig & Pleskac, 2010). Translated into the process of self-reporting one's risk preference, this implies that a single yet "strong" piece of information may outweigh multiple "weak" pieces of information. The strength of evidence might play a focal role in the process of rendering self-reports, given the observations of Griffin and Tversky (1992) as well as Kvam and Pleskac (2016) that people tend to focus on the strength of evidence rather than on the weight of evidence when rendering judgments based on external samples. To date it remains untested to what extent the strength of evidence of available information is relevant in the context of rendering self-reports of risk preference.

Third, a large body of research into serial-position effects suggests that people are highly sensitive to the order of information (e.g., Hertwig, Barron, Weber, & Erev, 2004; Hogarth & Einhorn, 1992; Yechiam & Busemeyer, 2005). For example, it has been observed that the endowment effect may at least in part result from order effects in the aggregation process of respondents' internal samples, as information retrieved in the beginning was more indicative of participants' judgments (E. J. Johnson et al., 2007). Whereas some research into the sequential aggregation of internal or external samples has found such primacy effects (e.g., E. J. Johnson et al., 2007; Weber et al., 2007), other research suggests the occurrence of recency effects (e.g., Barron &

Yechiam, 2009; Highhouse & Gallo, 1997; Hogarth & Einhorn, 1992). Hence, to the extent that people rely on multiple pieces of information when rendering self-reports, accounting for order effects may be important in the sense that information retrieved either at the beginning or at the end of the internal sampling process may be particularly influential.

Taken together, in contrast to research on decision making based on *external samples*, far less research exists on judgment formation based on *internal samples*—as are potentially drawn when rendering a self-report of one's risk preference. We aim to take a step towards closing this gap, by fostering a better understanding of the information-integration processes taking place when people render self-reports.

Aspect Listing: A Tool to Unpack Self-Reports

As information-integration processes typically remain hidden from direct observation, some research has employed the process-tracing method of *aspect listing* to gain a window into people's minds (E. J. Johnson et al., 2007; Weber et al., 2007; for a review see Schulte-Mecklenbeck et al., 2017). Specifically, this methodology entails prompting people to sequentially list their thoughts—typically referred to as *aspects*—that spontaneously cross their minds when responding to judgment or valuation questions. That is, by *not* prompting people to reflect on how they rendered a self-report *in hindsight*, this process-tracing method aims to avoid triggering any unnatural metacognitive processes, including potentially distorted post-hoc rationalizations (Nisbett & Wilson, 1977; but also see Hurlburt & Heavey, 2001).[1] Instead, and much like in research relying on think-aloud protocols, this method aims to trace the natural information-integration process "on the fly" (Ericsson & Simon, 1980).

Previous research adopting aspect listing has yielded several important insights, including into the cognitive processes underlying the endowment effect (e.g., E. J. Johnson et al., 2007), inter-temporal choice (e.g., Appelt, Hardisty, & Weber, 2011; Weber et al., 2007), the effect of attribute framing on choice (e.g., Hardisty, Johnson, & Weber, 2010), or domain-specificity in evolutionary content-domains (Jarecki & Wilke, 2018). Moreover, methods related to aspect listing ("thought-protocols" collecting information in a somewhat less structured way and after a judgment has already been provided) have also permitted several *qualitative insights* into judgment formation: For example, when rendering self-reports of life satisfaction (Schimmack, Diener, & Oishi, 2002) or risk preference (Arslan et al., 2020), people appear to rely mostly on personal experiences rather than on social comparisons (for details, see the introduction of study 1). Yet, contrary to the method of aspect listing these latter approaches do not readily permit the quantitative modeling of any information-integration processes, as the respective thought-protocols are typically not broken down into "atomic components" of evidence (e.g., the weight vs. strength of evidence).

---

[1] Aspect listing may be complemented by prompting respondents to provide additional ratings of the aspects they had previously listed (e.g., how strongly an aspect speaks in favor of or against a particular choice or judgment), and such additional ratings would thus classify as a metacognitive task (Greifeneder & Schwarz, 2014; Hurlburt & Heavey, 2001; Koriat, 2007; Koriat et al., 2000).

Overview and Research Aims

The goal of this article is to promote a better understanding of the psychology underlying people's self-reports of their risk preferences: Although people are routinely asked to provide such self-reports in scientific and applied contexts, the underlying information-integration processes and people's respective cognitive representations remain largely unknown.

To this end, study 1 implemented a cognitive modeling approach to account for people's self-reported risk preferences based on various quantitative dimensions of evidence—as extracted from the listed aspects. Specifically, to what extent does a *cognitive account* potentially outperform various sociodemographic variables in predicting interindividual differences in self-reported risk preferences (RQ1a)? And how influential are the three reviewed properties of evidence in people's information-integration processes (RQ1b)? Moreover, by analyzing the content of the listed aspects, study 1 also permitted obtaining a range of qualitative insights into the cognitive representations of people's risk preferences (RQ2).

Subsequently, study 2 aimed at testing a longitudinal hypothesis that logically follows from the assumption that people's self-reports of their risk preferences emerge from quantifiable information-integration processes and robust cognitive representations. Specifically, the high temporal stability of self-reported risk preference, as observed repeatedly in previous research (e.g., Frey et al., 2017), may originate from relatively stable cognitive representations of one's own behaviors and experiences. Therefore, in a retest study we examined the stability of the content of the listed aspects (RQ3a) and the stability of the listed aspects' strength of evidence (RQ3b), to test whether stability and change in any of these two dimensions are systematically associated with stability and change in self-reported risk preference (RQ4a and RQ4b).

In addressing these research aims, we attached great importance to adhering to transparent and reproducible scientific practices and thus published a preregistration including the full theoretical rationale, all data, and the analyses scripts at `https://osf.io/gndjw`.

Study 1

Information-integration processes have long been of central interest in the literature on judgment and decision making (e.g., Dawes & Corrigan, 1974; Gigerenzer & Goldstein, 1996; Hastie & Park, 1986; Payne et al., 1988), and a diverse set of modeling approaches has thus emerged in this regard. For instance, in the framework of the Brunswikian lens model the cognitive integration of external cues into a judgment has been modeled descriptively by means of simple linear models (Hammond & Stewart, 2001; Hastie & Dawes, 2001), which were also used to address normative questions concerning information integration in various judgment processes (e.g., the role of proper vs. improper linear models in decision making; Dawes, 1979). Another substantive body of research has focused on noncompensatory heuristics to examine information use and integration in the context of inferential choice (e.g., take-the-best, TTB; Gigerenzer & Brighton, 2009; Gigerenzer & Goldstein, 1996, 1999).

Finally, research into sequential information integration has developed sophisticated fractional-adjustment models to study, for instance, the role of serial position effects in information integration (e.g., Hogarth & Einhorn, 1992; Sutton & Barto, 1998).

To address RQ1 in study 1—that is, how well people's self-reported risk preferences can be quantitatively accounted for based on the listed aspects (RQ1a), and how influential different properties of these aspects are in people's self-reports (RQ1b)—we built on these different strands of research and implemented a twofold-approach. First, we directly sampled a set of models from the literature, aimed at covering a large model space to thus incorporate models that account for (different combinations of) the strength of evidence, the weight of evidence, and the order of evidence. As reviewed above, these dimensions constitute three key properties of evidence that people may rely on when rendering self-reports. Our approach followed a proof-of-concept provided by Jarecki and Wilke (2018), who used cognitive process models to study risk taking in different evolutionary domains. Yet, our approach was different in the sense that it focused on general risk preference, modeled continuous self-reports (as opposed to hypothetical binary choices), and importantly, took into account the listed aspects' strength of evidence—a property that may be highly relevant during information integration according to previous observations (Griffin & Tversky, 1992; Kvam & Pleskac, 2016). Second, as some of the different models turned out to yield similar predictions when applied to the empirical data of study 1 we also implemented a set of Bayesian ordinal regression models using the listed aspects' weight of evidence and strength of evidence as direct predictors of self-reported risk preference—to thus facilitate a direct comparison between the roles of these two properties of evidence.

Beyond analyzing quantitative aspects of the information-integration processes (i.e., RQ1a and RQ1b), the method of aspect listing also permitted conducting a series of more qualitative analyses, which allowed insight into people's cognitive representations of their risk preferences (RQ2). Specifically, these analyses characterized the content and sources of the aspects people rely on during information integration, such as whether people predominantly tap into personal experiences or social comparisons to render their self-reports (e.g., Arslan et al., 2020; Schimmack et al., 2002), or how frequently people typically experience in daily life what they consider as aspects during judgment formation. Previous research along these lines, which has prompted respondents to *explain* their previously stated risk preferences, found that people mainly considered risks that they had personally taken, which were rather voluntary, had known and controllable consequences, and were old and familiar (Arslan et al., 2020). Our approach promised to corroborate and extend these findings, as we prompted participants to *concurrently* list the aspects that crossed their minds *during judgment formation* (i.e., as opposed to after already having provided a response, which in principle could lead to distorted reports; Nisbett & Wilson, 1977). Moreover, as the approach of aspect listing taps into people's cognitive representations in a semi-structured way (i.e., collecting aspects one by one), it is possible to examine, for instance, the content and sources separately for aspects that speak either pro or contra risk taking (i.e., pro-aspects and contra-aspects, respectively). In addressing RQ2, we again relied on a two-fold approach: We first analyzed the cognitive representations based on respondents' own ratings of their aspects. Second, we also relied

on the evaluations of a subset of 300 aspects as provided by external raters. The latter evaluations rendered possible further insight concerning the content of the aspects (e.g., classification to various content domains) as well as an external validation of the aspects' strength of evidence.

Methods

We collected data from 250 participants via Amazon MTurk (115 females; mean age: 37.4 years; range: 18 – 73 years; mean number of years of education: 15.2; modal income: 1,000 - 2,000 USD per month). To ensure a high data quality, only MTurkers with an approval rate of at least 95% and who had completed at least 500 HITs (i.e., human intelligence tasks) on Amazon MTurk were eligible to participate (Peer, Vosgerau, & Acquisti, 2014; see also Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013; Paolacci, Chandler, & Ipeirotis, 2010). Moreover, participants had to pass two attention check questions and provide ratings of at least 25 out of 100 on questions asking how focused they were and how much effort they put into the study. Data were collected in 2019. Study completion on average took 6 minutes, for which participants were reimbursed with 0.85 USD. Both studies were approved by the ethics committee of the Faculty of Psychology of the University of Basel (#023-18-1).

According to a prior model recovery analysis (see preregistration; cf. Gluth & Jarecki, 2019), a sample of 250 participants was sufficiently large for the separate models to be recovered with high recovery rates, except for two models which were thus excluded from the model space. This sample size is also sufficient to detect small to medium effects in a frequentist framework (f^2 of .03 with a power of $1 - \beta = .80$; calculated using G^*Power 3.1, Faul, Erdfelder, Buchner, & Lang, 2009) for the most complex regression model involving three predictors (i.e., query theory, see below). Note, however, that we conducted all analyses in a Bayesian framework and we thus report 95% credible intervals (95% CIs) rather than p-values (unlike in a frequentist framework, the 95% CI indicates the range that contains the population parameter with a probability of 95%).

All analyses were performed using *R* (R Core Team, 2020). We used the *rstanarm* and *brms* packages for the regression analyses (Bürkner, 2017; Goodrich, Gabry, Ali, & Brilleman, 2018) and implemented the default priors as provided by these packages (see Supplemental Material, SM, section 4.3).

Procedure. After reading general instructions, participants provided informed consent and sociodemographic information. They were then shown the general risk item of the SOEP ("Are you generally a person who is willing to take risks or do you try to avoid taking risks?"; e.g., Dohmen et al., 2011). Yet, prior to actually providing an answer participants were prompted to think of (and list) all reasons that crossed their minds *while* coming up with an answer (the exact wording is available on `https://osf.io/gndjw`). Specifically, participants had to report at least one aspect and were asked to continue reporting aspects until they could not think of any further aspects. Once done with this task, participants provided their rating to the SOEP general risk item on a scale ranging from 0 to 10. The procedure of first implementing

the aspect listing followed the original protocol (E. J. Johnson et al., 2007; Weber et al., 2007); we specifically pretested potential order effects (i.e., self-reported risk preference first vs. aspect listing first) in a dedicated pilot study (see preregistration), which revealed no credible mean differences of self-reported risk preferences in the two examined orders. Finally, participants were sequentially presented with the aspects that they had previously listed (in randomized order), and were prompted to evaluate these aspects on a series of dimensions (i.e., including the aspects' strength of evidence as well as dimensions tapping the content and sources of the aspects; see Table 1 and the respective sections below).

Table 1

Assessment of the Aspects' Properties in Study 1

Property/Source	Derived From	Operationalization
Properties used to model self-reported risk preferences		
Strength of evidence	Rating by participant	"How strongly does your description above support that you seek risks vs. that you avoid risks?" (ranging from 50, labeled "strong support for risk seeking," to -50, labeled "strong support for risk avoidance")
Weight of evidence	Binarized strength of evidence	Number of pro-aspects (strength of evidence > 0) and number of contra-aspects (strength of evidence < 0). Neutral aspects (strength of evidence $= 0$) were ignored in the analysis.
Order of evidence	Sequence of the listed aspects	Depending on the models implementing serial-position effects (see QT and VUM).
Properties to explore sources and content of evidence		
Personal experience	Rating by participant	"Does your description above include a personal experience?"
Social comparison	Rating by participant	"Does your description above include a comparison with another person?"
Frequency in daily life	Rating by participant	"How frequently do you normally experience or do what you described above?"
Active choice vs. passive experience	Rating by participant	"Does what you described above include something you actively chose or something you passively experienced?"
Controllability	Rating by participant	"Are the outcomes and consequences of what you described above controllable or uncontrollable for you?"
Sentiment	Wording of aspects	A score of the sentiment (positive or negative) of an aspect.

Note: The ratings concerning personal experience, social comparison, active choice, and controllability were binary ("Yes, it includes a personal experience" vs. "No, it does not include a personal experience" for the personal experience rating; "Yes, it includes a comparison with another person" vs. "No, it does not include a comparison with another person" for the rating of social comparison; "Rather controllable" vs. "Rather uncontrollable" for the rating of controllability; and "Active choice" vs. "Passive experience" for the rating concerning active choice vs. passive experience). Frequency was assessed categorically, with the categories "Once per day", "Once per week", "Once per month", "Once per year", "Less than once per year, but on a regular basis", and "Less than once per year, only once or a few times so far". In all ratings on the sources and content, a "Not applicable" option could be selected.

Procedure and analyses concerning RQ1a and RQ1b. To formally model the information-integration processes underlying people's self-reports (RQ1a and RQ1b), we implemented the following steps.

Operationalization of the aspects' properties of evidence. We operationalized the three quantitative properties of evidence reviewed above as follows (see Table 1): First, we operationalized the strength of evidence as participants' ratings of how strongly each aspect supports risk-avoidance or risk-seeking, ranging from -50 to 50. Second, to determine the weight of evidence, we classified the aspects—based on the rated strength-of evidence—as either pro-aspects (strength of evidence > 0) or contra-aspects (strength of evidence < 0) and then counted the number of pro- and contra-aspects for each participant. Third and finally, the order of evidence naturally followed from the sequence by which participants listed the aspects.

Model space and model selection criteria. We initially implemented six separate models (for a detailed description of all models, see SM section 4.1) to cover various combinations of the three properties of evidence as reviewed above. Specifically, the EXT model (inspired by the TTB heuristic; Gigerenzer & Goldstein, 1996, 1999) used the most extreme strength of evidence (i.e., the one the furthest away from the center of the scale) as predictor; the FIRST model used the strength of evidence of the aspect listed first in the sequence as predictor (for related lexicographic models such as take-the-first; see Jarecki & Wilke, 2018; J. G. Johnson & Raab, 2003); and the LAST model used the strength of evidence of the aspect listed last in the sequence as predictor. These three models were non-compensatory models; the remaining three models were compensatory models. Specifically, the SUM model (a weighted additive model; see Payne et al., 1988) used the sum of the strength of evidence of all aspects listed by a participant as predictor; query theory (QT; E. J. Johnson et al., 2007; Weber et al., 2007) was implemented as a linear model with the weight of evidence (the number of pro-aspects and the number of contra-aspects separately) and the order of evidence as predictors; and finally, the value updating model (VUM; an instance of a fractional-adjustment model; Hertwig et al., 2006; Hogarth & Einhorn, 1992) implemented a weighted average of the strength of evidence as predictor, rendering possible the capture both of primacy and recency effects.

To enable a fair model comparison—accounting for the fact that some models (i.e., QT and VUM) had free parameters whereas others did not—we purely focused on predictive accuracy (i.e., out-of-sample prediction). Thereby, adjustable parameters only provide an advantage for a model if they actually help explain systematic variance (e.g., Yarkoni & Westfall, 2017). To this end, we employed a five-fold cross-validation approach. That is, we partitioned the data in five subsets (folds) and used four folds to fit the free parameters (i.e., in the case QT and VUM) and predicted the fifth (hold-out) fold with the obtained parameter estimate. Given our data structure with one response per participant, all parameters were estimated across participants. This procedure was repeated for all models until each of the five folds was predicted once by every model. We then determined the average (i.e., across the independent hold-out samples) Spearman rank correlations (i.e., r_s) between the model predictions and the self-reported risk preferences.

Based on a prior model recovery analysis (see preregistration), the six initial models were expected to yield somewhat correlated yet sufficiently distinguishable predictions. Ultimately, however, in study 1 the models ended up making relatively similar predictions, given the average of 3.4 aspects that participants listed (note that this number is in line with previous studies, e.g., Jarecki & Wilke, 2018; E. J. Johnson et al., 2007; Weber et al., 2007), and given that participants tended to list either only pro-aspects or only contra-aspects (which also made it difficult to systematically study order effects). To illustrate, FIRST and LAST resulted in very similar model predictions, and a parameter recovery analysis for the VUM indicated that different values for the weighting parameter (i.e., capturing recency or primacy effects) resulted in very similar model predictions (see SM section 4.5).

We thus also pursued a complementary approach as a robustness check. Specifically, we employed two Bayesian ordinal regression models, using the aspects' strength of evidence and weight of evidence (i.e., averaged per participant; when averaging the weight of evidence each pro-aspect was given the value 1, and each contra-aspect was given the value -1), respectively, to predict self-reported risk preferences (see Table 1). We relied on multiple indices to compare these models: First, we compared their expected log predictive density (ELPD)—a statistic that provides an estimate of the to-be-expected out-of-sample predictive performance—based on the leave-one-out information criterion (LOOIC), which is similar to the Akaike information criterion (AIC) but better suited for Bayesian model comparisons, as it can account for the implemented priors (Vehtari, Gelman, & Gabry, 2017). Moreover, we compared the models' accuracies, their chance corrected accuracies (Cohen's κ), as well as how often their (correct) predictions coincided in a tournament approach (see Broomell, Budescu, & Por, 2011).

Reference models. To compare the described models against a baseline, we also implemented three Bayesian ordinal regression models that inferred participants' self-reported risk preferences based on up to five sociodemographic predictors. The first model included age as the sole predictor, the second model included sex as the sole predictor, and the third model included age, sex, years of education, income, and employment status as predictors. These variables have been suggested to be systematically associated with individual differences in risk preference, and in the case of age (e.g., Mamerow, Frey, & Mata, 2016; Mata, Josef, & Hertwig, 2016) and sex (e.g., Byrnes, Miller, & Schafer, 1999), these associations were found to be particularly robust (for an overview, see Frey, Richter, Schupp, Hertwig, & Mata, 2020). To compare these reference models with the other two ordinal regression models, we included them in the tournament approach described above, and additionally relied on the LOOIC-based ELPD. Finally, to compare the reference models with the initial set of models described above, we also report Spearman correlations between the model predictions and participants' self-reported risk preferences.

Procedure and analyses concerning RQ2. To examine the content and sources of the aspects people rely on during judgment formation (RQ2), we implemented the following steps.

Participants' own ratings of their aspects. At the end of the study, participants provided ratings of each aspect they had previously listed, concerning

(a) how strongly the aspect supported risk seeking versus risk avoidance (i.e., the strength of evidence used in the modeling analysis; see above), (b) whether the aspect included a previous personal experience (see Arslan et al., 2020), (c) whether the aspect included a comparison with another person (see Arslan et al., 2020; Schimmack et al., 2002), (d) how often participants typically experience in their daily lives what they described in the aspect (i.e., relatively common or rather rare but potentially high-stake events; see Hertwig et al., 2004), (e) whether the aspect referred to an active choice or a passive experience (i.e., voluntary or involuntary exposure to risks; Fischhoff, Slovic, Lichtenstein, Read, & Combs, 1978), and (f) whether the aspect involved something controllable or uncontrollable (see Arslan et al., 2020; Fischhoff et al., 1978; Maccrimmon & Wehrung, 1985). Table 1 provides an overview and a detailed description of the items used.

For each of these dimensions, we provide the distributions of participants' ratings, separately for pro- and contra-aspects, and report post-hoc mixed-effects models to explore any systematic differences between pro- and contra-aspects. Specifically, we ran generalized linear mixed-effects models predicting the various ratings and using the aspects' direction (pro or contra) as dummy coded predictors, using by-subjects random slopes and intercepts (Barr, Levy, Scheepers, & Tily, 2013). We also quantified the differences in the sentiment for pro- and contra-aspects (see SM sections 2 and 4.2).

External ratings of a subset of 300 aspects. For 300 randomly selected aspects (i.e., about one third of the 857 aspects listed in study 1), we also collected external ratings from three independent raters (i.e., the first author and two research assistants; using a majority rule to integrate the three ratings; see SM section 5.5 for further methodological details).

First, the raters inferred the listed aspects' strength of evidence, to thus provide an independent validation of participants' own ratings. To this end, we provided the same scale as participants used to evaluate their own aspects.

Second, the raters assessed a range of additional properties that were not assessed by participants themselves. These properties stem from five risk categories and have been suggested to be important drivers of and motives underlying risk-taking behaviors, covering both stable dispositions (i.e., traits) as well as situational characteristics (i.e., state variables), namely: (a) outcome-related properties (e.g., the magnitude of the positive outcomes; Kahneman & Tversky, 1979; Sitkin & Pablo, 1992), (b) goal/state-related properties (e.g., whether the goal was to keep or improve one's status quo; e.g., Lopes, 1984; Mishra, Barclay, & Sparks, 2017), (c) properties related to cultural roles and personality (e.g., whether a social norm or one's personality was mentioned; Nicholson, Soane, Fenton-O'Creevy, & Willman, 2005; Sitkin & Pablo, 1992), affect-related properties (e.g., whether a feeling of fear or thrill was mentioned; Lerner, Gonzalez, Small, & Fischhoff, 2003; Loewenstein, Weber, Hsee, & Welch, 2001; Zuckerman, 2002), and (d) properties related to life-history (e.g., whether one's age or children were mentioned; e.g., Wang, Kruger, & Wilke, 2009). Finally, we used an *other* category to classify whether an aspect just relativized (e.g., "that depends on the situation"), or only contained semantically invalid sequences of letters. Please see SM section 5.5 for the complete list of properties along with some

key references and the full description of the rating procedure.

Third, the raters inferred the life domains to which the listed aspects supposedly belong to. To this end we provided the domains as suggested in one of the most popular domain-specific risk-taking questionnaires (DOSPERT; Blais & Weber, 2006; Frey, Duncan, & Weber, 2020; Weber, Blais, & Betz, 2002), those suggested in the SOEP (e.g., Dohmen et al., 2011), as well as those of the evolutionary risk scale (ERS; Wilke et al., 2014)—overall resulting in 19 different domains (see SM section 5.5).

Results

In line with previous observations, the self-reports of the majority of participants (57%) indicated risk-aversion (i.e., most participants provided a rating of lower than five on the scale ranging from 0 to 10), with an average rating of $M = 4.2$. The majority of participants (81%) listed between one and four aspects ($M = 3.4$; range: 1 – 12). Matching participants' overall tendency for risk-aversion, the majority of these aspects were contra-aspects (61%). Moreover, most participants (82%) only listed either contra-aspects or pro-aspects, directionally matching their risk preference (i.e., risk-seeking vs. risk-averse). Participants' ratings of their aspects' strength of evidence were relatively consistent within participants, with an intra-class correlation of .76.[2]

Our validation of the aspects' strength of evidence using external raters showed a high degree of agreement: The strength of evidence as assessed by the external raters (i.e., average across the three raters) and the strengths of evidence as indicated by participants themselves correlated with $r_s = .82$. Moreover, in 93% of the cases the three raters classified the listed aspects correctly (i.e., in line with participants' own judgments) as pro- or contra-aspects.

RQ1: Modeling self-reported risk preferences. As outlined above, we followed a two-fold approach to modeling self-reported risk preferences. First, we compared six separate models directly sampled from the literature on judgment and decision making. These models were capable of predicting self-reported risk preference well, with r_s ranging from .78 to .90; note that these values resulted from out-of-sample predictions using the independent hold-out sets. Specifically, the correlations between model predictions and actual self-reports were $r_s = .90$ (VUM), $r_s = .83$ (EXT), $r_s = .78$ (QT), $r_s = .82$ (FIRST), $r_s = .84$ (LAST), and $r_s = .82$ (SUM). Moreover, these models clearly outperformed the three reference models, with correlations between the predictions of the latter and the self-reports ranging from $r_s = .14$ to $r_s = .31$ (see Figure 1).

Second, we compared a set of ordinal regression models using the strength of evidence (SoE) and the weight of evidence (WoE) as direct predictors. Corroborating the results reported above, both models performed well, with $r_s = .90$ (SoE) and $r_s = .85$ (WoE). Moreover, the model including the strength of evidence as predictor outperformed the model including the weight of evidence as predictor by eight

[2]This analysis was run with an intercept-only model, with by-subjects random intercepts predicting the aspects' strength of evidence.

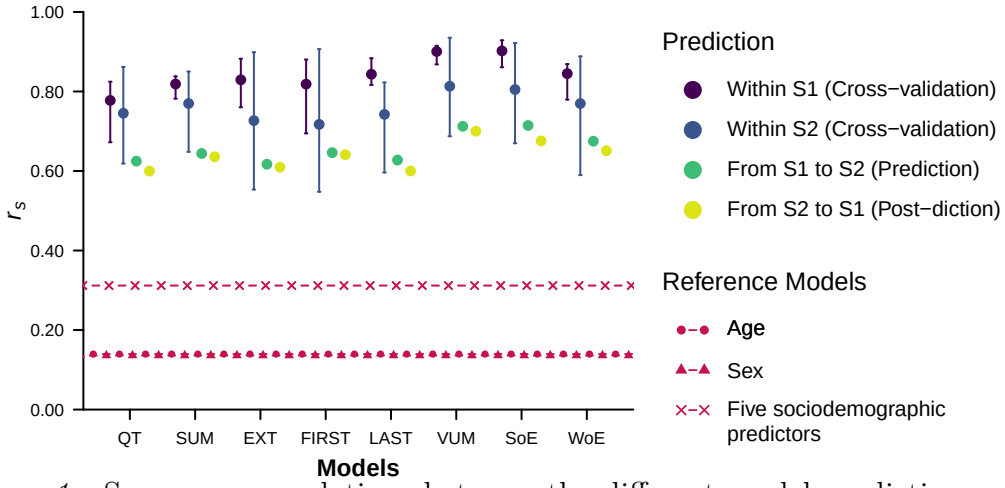

Figure 1. Spearman correlations between the different model predictions and self-reported risk preference. QT = Query theory; SUM = Sum of evidence; EXT = Most extreme evidence; FIRST = First aspect's evidence; LAST = Last aspect's evidence; VUM = Value updating model. SoE = Ordinal regression model with the average strength of evidence per participant as predictor. WoE = Ordinal regression model with the average weight of evidence per participant as predictor. Whiskers depict the range of r_s in the five folds of the cross-validation within studies. For the pre-/post-diction across studies, the models only used the aspects participants listed in one study to pre-/post-dict their risk preferences in the other study. "Five sociodemographic predictors" = Reference model using age, sex, years of education, income, and employment status as predictors. All reference models were implemented separately for study 1 and study 2 and their respective r_ss averaged for this plot.

percentage points of *correct predictions* (see Table 2). Also, there was robust evidence that the strength of evidence was a more important predictor than the weight of evidence according to the direct model comparison based on the two models' to-be-expected out-of-sample predictive performance (i.e., LOOIC-based ELPDs; see Table 2).

Following the tournament approach proposed by Broomell et al. (2011), we also gauged the proportion of identical model predictions of the five ordinal regression models (i.e., the two models using the strength and weight of evidence as predictors, and the three reference models). While some models resulted in highly similar predictions (i.e., the reference models including only age or sex as predictors made identical predictions in 97% of the cases), the two models using the different properties of evidence as predictors were sufficiently distinguishable (see Table S2 and Figure S5; see also Table 2).

Finally, again in line with the comparison of the models reported above, both tested properties of evidence proved to be better predictors than any of the reference models that used sociodemographic predictors. Specifically, the strength of evidence model outperformed the best reference model by 18 percentage points, and the weight of evidence model outperformed the best reference model by ten percentage points (see Table 2).

Table 2
Goodness of Fit Indicators of the Different Ordinal Regression Models.

Model	Accuracy	κ	Distinct Predictions	ELPD
Study 1				
SoE	.42	.37	7	0 [0, 0]
WoE	.34	.28	5	-57.4 [-75.0, -39.8]
5 soc. dem. pred.	.24	.16	3	-194.6 [-219.6, -169.6]
Sex	.19	.11	2	-194.7 [-218.3, -171.1]
Age	.20	.12	3	-195.8 [-220.0, -171.6]
Study 2				
SoE	.33	.27	7	0 [0, 0]
WoE	.31	.24	5	-21.4 [-34.4, -8.4]
5 soc. dem. pred.	.23	.16	4	-84.8 [-111.6, -58.6]
Sex	.20	.12	3	-82.4 [-109.4, -55.4]
Age	.21	.13	1	-82.6 [-110.2, -55.0]

Note: Results are based on ordinal regression models (not cross-validated). Accuracy = The proportion of correctly predicted categories (ratings between 0 and 10). κ = Cohen's kappa, with a chance level of 1/11. Distinct Predictions = The number of distinct/unique predictions made by a model (all numbers from 0 to 10 occurred in the empirical data). ELPD = Estimate of the leave-one-out information criterion based expected log predictive density for a new dataset, relative to the best model (i.e., SoE)—where lower numbers indicate worse model fit. ± 2 standard errors interval are given in brackets. SoE = Mean strength of evidence per participant as predictor. WoE = Mean weight of evidence per participant as predictor. 5 soc. dem. pred. = Age, sex, years of education, income, and employment status as predictors. Sex = Sex as predictor. Age = Age as predictor.

RQ2: Sources and content of the listed aspects. Our analyses of people's cognitive representations of their risk preferences (see Figure 2) indicated that most participants retrieved personal experiences (and less so social comparisons) when rendering their self-reports (more so for pro- than contra-aspects: b = 1.87, 95% CI: [0.82, 3.17]). Furthermore, the listed aspects involved mostly active choices rather than passive experiences (more so for pro- than contra-aspects: b = 1.44, 95% CI: [0.81, 2.22]), and situations with rather controllable outcomes (no credible differences between pro- and contra-aspects: b = 0.24, 95% CI: [-0.52, 1.08]). Across the listed aspects, participants' answers to these questions were quite consistent; that is, most participants rated their respective aspects similarly on a given question. Furthermore, the listed aspects were typically not rare situations or experiences, but frequent encounters in participants' daily lives (i.e., the categories once per day, once per week, and once per month made up for 80.4% and 74.8% of all pro- and contra-aspects, respectively). Finally, most aspects had a negative sentiment (see SM section 2), but the pro-aspects less so than contra-aspects ($M_{pro-aspects}$ = -0.52; $M_{contra-aspects}$ = -1.12; b = 0.60, 95% CI: [0.40, 0.78]).

As Figure 3 illustrates, positive emotions and feelings as reflected by the words *fun* or *enjoy* often occurred in pro-aspects, along with words describing positive out-

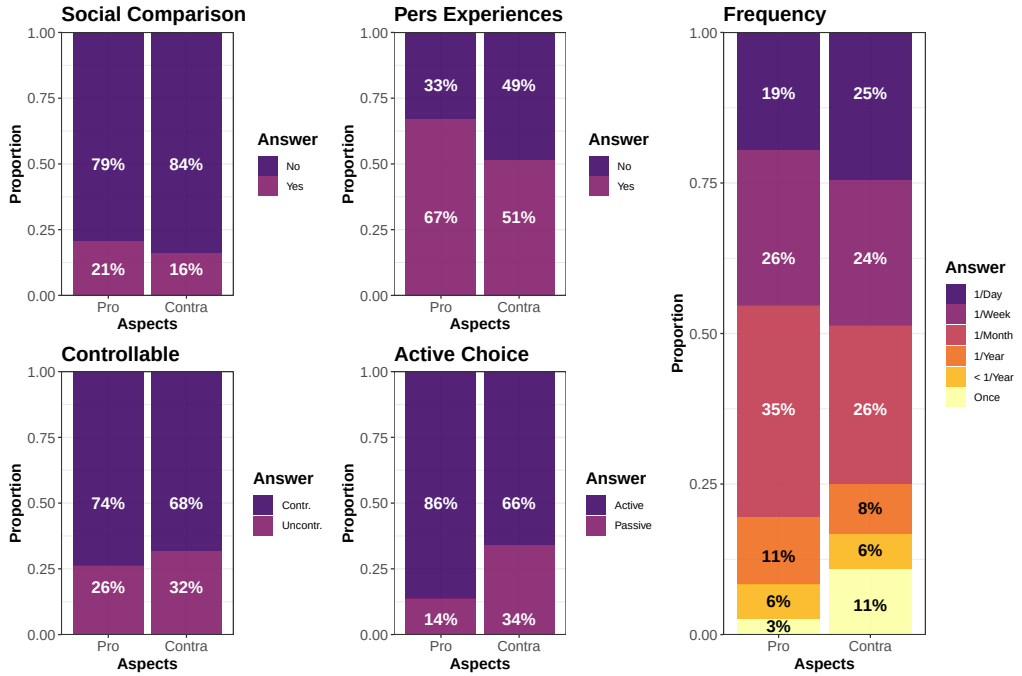

Figure 2. Distributions of the sources and content of the listed aspects from study 1 across all participants and aspects.

comes such as *reward*, *gain*, or *benefit*. The picture looked substantially different for contra-aspects, where *lose*, *money*, or *safe* where very prominent mentions, along with negative emotions or feelings expressed by words such as *hurt*, *afraid*, or *worry*.

Our additional analyses using external ratings of the listed aspects showed that participants mostly retrieved domain-general statements (79.8%), and if domain-specific statements were retrieved, these were mostly in the domains of health/safety (9.8%), financial (7.1%), social (4.0%), occupation (3.4%), recreational (2.7%), and kinship (2.0%).[3] In the SM (section 5.5) we report an analysis showing that the domains put forth by two established domain-specific risk-taking scales could be recovered well—that is, in 13 out of the 15 distinct domains suggested by these two scales, more than half of the respective items were correctly recovered, and in eight of the 15 domains all items were correctly recovered. Regarding the potential drivers and motives underlying risk taking, we found that participants mostly considered the valence of the potential outcomes (i.e., positive outcomes in the pro-aspects, but also often in combination with negative outcomes—i.e., indication of a risk-return trade-off; and more negative outcomes in the contra-aspects). Moreover, participants often mentioned their positive (in the case of pro-aspects) and negative feelings (in the case of contra-aspects) towards taking risks. Finally, in pro-aspects participants often adopted an opportunity focus, aimed at improving their status quo, while in contra-aspects, participants often adopted a safety focus, aimed at keeping their status quo (see also Figure S9).

[3]The external raters could select multiple domains per aspect, which is why these numbers do not add up to 100%.

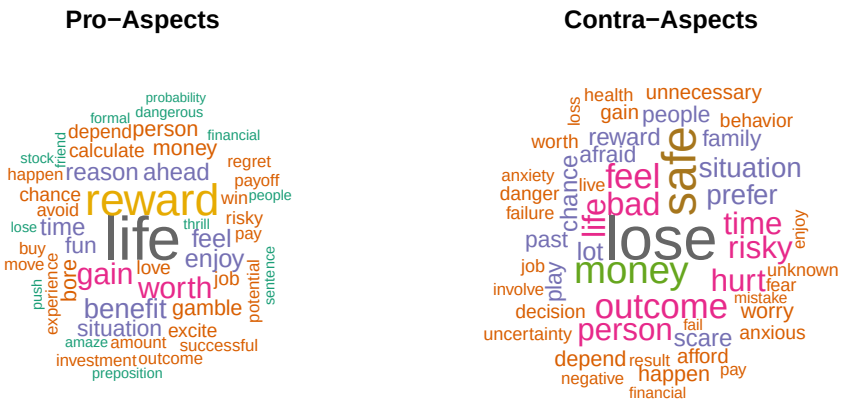

Figure 3. Word cloud of the most frequently occurring words separate for pro- and contra-aspects. Larger fonts indicate higher frequencies of occurrence of the respective words in the aspects listed in study 1. The words *risk*, *take*, and *avoid* were excluded to increase the visibility of the other words that occurred less frequently than these words that were also part of the formulated question and thus were often repeated in the aspects (e.g., "I avoid taking risks because..."). The words listed in the word cloud of pro-aspects had a clearly positive average sentiment (M = 7.07), and those listed in the word cloud of contra-aspects had a clearly negative average sentiment (M = -3.60)

Discussion

The listed aspects—and more precisely, different properties of evidence thereof, particularly the strength of evidence—turned out to be highly predictive of participants' self-reported risk preferences: Overall the cognitive modeling approach performed substantially better than using a series of sociodemographic indicators as predictors (i.e., reference models), which suggests that people recruit systematic information-integration processes when rendering self-reports of their risk preferences. Crucially, external ratings of the aspects' strength of evidence were closely aligned with participants' own ratings, thus making it unlikely that the high predictive power of this property of evidence simply arose due to the close temporal proximity between the respective ratings (note that we also conducted an independent cross-study analysis as a further robustness check in this regard, see below).

Although we faced some constraints when comparing the six initial models sampled from the literature—particularly concerning the role of the order of evidence—a direct evaluation of the relative importance of the aspects' strength of evidence and weight of evidence yielded clear and corroborating results. Furthermore, given the structure of our data it may seem somewhat surprising that the non-compensatory models were outperformed by the compensatory models (see Gigerenzer & Gaissmaier, 2011; Gigerenzer & Goldstein, 1996), yet it is important to keep in mind that these differences in model performance were small.

Participants' cognitive representations of their risk preferences proved to rest mostly on situations that involved active choices rather than passive experiences, suggesting that most people may think of risk taking as an explicit decision. The mostly domain-general information retrieved by participants tended to focus on the valence of the outcomes, often referring to explicit trade-offs in line with a risk-return framework (Weber et al., 2002; Weber & Milliman, 1997). In pro-aspects, participants often expressed an opportunity focus, whereas in contra-aspects they often expressed a safety focus. Moreover, in line with some conceptualizations of risks (e.g., Bell, 1982; Loewenstein et al., 2001; Loomes & Sugden, 1982; Mellers, Schwartz, Ho, & Ritov, 1997), many aspects mentioned positive or negative feelings.

Taken together, the results of study 1 suggest that when people are prompted to report their own risk preferences, they may retrieve information from memory and evaluate *how strongly* multiple pieces of information support a specific judgment (i.e., strength of evidence). Thus, on a quantitative level the information-integration processes in the context of evaluating one's own risk preferences appear to share similarities with those of evaluating external objects (Griffin & Tversky, 1992; Kvam & Pleskac, 2016).

Study 2

In study 2 we tested a longitudinal hypothesis that logically follows from the basic assumption that people's self-reported risk preferences are robustly rooted in their cognitive representations of idiosyncratic experiences and behaviors. Specifically, in RQ3 we tested whether two particular dimensions of the retrieved information show stability across time—that is, from study 1 to study 2—namely, (a) whether people retrieve the identical aspects (aspect stability) and (b) whether the listed aspects have, on average, a similar strength of evidence (evidence stability). To illustrate, to the extent that people sample aspects from a large pool of idiosyncratic experiences, they may not necessarily retrieve the exact same aspects at different occasions (e.g., because different contexts may prime the retrieval of a particular type or class of aspects)—yet this naturally does not preclude the possibility that the retrieved aspects still suggest a similar degree of risk preference. Consequently, RQ4 examined whether the stability of self-reported risk preferences directly hinges on aspect stability or on evidence stability.

Methods

Of the 250 participants in study 1, 164 accepted an invitation to complete a retest study after an interval of one month. Of these participants, 150 passed all quality checks and their data were used for the subsequent analyses (72 females; mean age: 39.07; range: 19 – 70 years; mean number of years of education completed: 15.44; modal income: 2,000 - 3,000 USD per month). We deviated from our preregistered analysis plan on four minor points (see SM section 3).

The participants who completed both studies did not differ credibly from participants who only completed study 1 in terms of their self-reported risk preferences,

average strength of evidence of the listed aspects, average sentiment of the listed aspects, years of education, or the proportion of females (see SM section 5.6); however, the former participants tended to be slightly older ($b = 2.90$, 95% CI [0.16, 5.40]) and on average listed slightly more aspects ($b = 0.73$, 95% CI [0.26, 1.15]). In sum, if at all there were only very weak indications for systematic selection effects.

Procedure. The design of study 2 was equivalent to that of study 1, with the exception that we added two questions at the end of the study. Specifically, we asked participants how well they could remember the aspects they had listed in study 1, as well as concerning their intuition of how similar their listed aspects were across studies. Both of these ratings were provided on a scale ranging from 0 to 100. Participants again received compensation of 0.85 USD for their participation.

External similarity ratings to gauge aspect stability. To examine aspect stability, we first obtained similarity ratings for the listed aspects. To this end, we asked 63 independent raters (recruited via Amazon MTurk) to judge the similarities of all possible pairs of aspects that were listed by each participant across and within the two studies. Pairs of aspects were partitioned into packages of about 200, and for each package three raters were asked to provide their judgments using a Likert scale ranging from 0 to 5 (i.e., each rater rated a total of around 200 aspect pairs, one pair at a time). To gauge the inter-rater agreement we calculated Kendall's coefficient of concordance (W; Kendall, 1948) for each triplet of raters who evaluated the similarities of the same aspects ($M_W = .56$, range = .38 - .77).

We denoted two aspects to be "equivalent" using a very conservative cutoff of five (i.e., mean similarity rating across the three raters, implying that all raters had to provide the highest rating). To obtain the proportion of equivalent aspects we divided the number of equivalent aspects by the maximal number of aspects that could be equivalent; across studies, the maximally possible number of equivalent aspects is equal to the smaller number of aspects listed in study 1 and study 2. As a robustness check, we also used additional ways to aggregate similarity ratings in the analyses concerning aspect stability (SM section 5.4).

Statistical analysis. To quantify the relation between aspect stability and the stability of the self-reported risk preferences (RQ4a), we used a gamma regression model with a log link function. This allowed us to account for the skewness in the absolute difference scores of the self-reported risk preferences. For the robustness test with the average similarity rating as predictor, we again used a gamma regression model with a log link function.

To quantify the relation between evidence stability and the stability of the self-reported risk preferences (RQ4b), we used a linear regression model with both the evidence stability and the self-reported risk preferences scaled for better interpretability. In contrast to the relation between overlaps and change in the self-reported risk preferences in RQ4a, the variables involved in RQ4b—that is, the change in the aggregated strength of evidence and the change in the self-reported risk preferences—allow for testing a directional relationship. Therefore, we did not use the absolute differences but the directional difference scores of the variables between study 1 and study 2. We again used the default priors implemented in *rstanarm*.

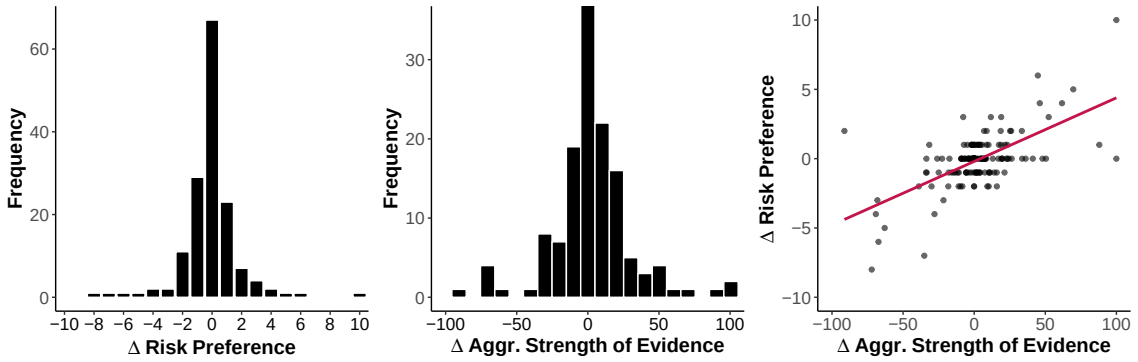

Figure 4. Stability of self-reported risk preference (first panel) and stability of the aspects' strength of evidence (second panel). The histograms show the distributions of within-subject differences between study 1 and study 2. The relation between changes in the aggregated strength of evidence and changes in self-reported risk preference is shown in the third panel ($r_s = .45$).

Results

Just as in study 1, the self-reports of the majority of participants (60%) indicated risk-aversion, with an average rating of $M = 3.81$. The majority of participants (78%) again listed between one and four aspects ($M = 3.6$; range: 1 – 13), and within participants the number of listed aspects was quite similar from study 1 to study 2 ($r_s = .54$). On average, participants indicated that they did not actively remember the aspects they had listed in study 1 ($M = 21.56$, $SD = 24.89$; on a scale from 0 to 100). Nevertheless, participants appeared to have an intuition that the aspects they had listed in study 1 were rather similar to those they had just listed in study 2 ($M = 65.42$, $SD = 22.05$; again on a scale from 0 to 100). The strength of evidence of the listed aspects again proved to be the most important predictor of participants' self-reported risk preferences (see Figure 1).

Finally, in line with previous observations (e.g., Frey et al., 2017; Mata et al., 2018) participants' self-reported risk preferences were highly stable at a one-month interval ($r_s = .80$). The first panel of Figure 4 depicts the distribution of within-subject differences, which is clearly centered on zero.

RQ3: Aspect stability and evidence stability. We examined aspect stability by determining the proportion of equivalent aspects across studies (see methods section). With the strict criterion imposed for classifying aspects as equivalent, aspect stability was relatively low: Only every twentieth aspect pair (i.e,. a proportion of .05) fulfilled the criterion of equivalence.

Yet, the picture was substantially different for evidence stability: Specifically, the strength of evidence (aggregated over all aspects listed by each participant)[4]

[4]In line with our preregistered analysis plan, we used the value updating model to aggregate evidence stability across the aspects listed by each participant, because it was the best-performing of the original models in both studies. Yet, using the arithmetic mean to aggregate the strength of evidence yielded an equivalent result ($r_s = .67$).

remained highly stable across time (r_s = .68), as can be seen in the second panel of Figure 4. A Bayesian paired t-test corroborated that there was no credible difference between the aggregated strength of evidence of a participant's aspects listed in the two studies (Δ_M = -1.39, 95% CI [-3.48, 0.64]).

RQ4: Relationship of the stability of self-reported risk preference with aspect stability and evidence stability. Aspect stability was not credibly associated with the stability of self-reported risk preference (b = -1.03, 95% CI [-2.90, 1.59]; r_s = -.12) nor did a robustness test (i.e., using the average similarity ratings instead of the proportion of equivalent aspects) indicate a credible association between aspect stability and the stability of self-reported risk preference (b = -0.43, 95% CI [-0.95, 0.10]; r_s = -.20).

Conversely, and as can be seen in the third panel of Figure 4, evidence stability across the two studies was credibly and strongly associated with the stability of self-reported risk preference (β = 0.63, 95% CI [0.50, 0.75]; r_s = .45).

Discussion

Study 2 corroborated the results obtained in study 1, and replicated previous observations of a high temporal stability of self-reported risk preference (e.g., Frey et al., 2017; Mata et al., 2018). More importantly, our analyses revealed that within participants the listed aspects' average strength of evidence remained highly stable across the two studies; that is, although participants did not necessarily list the exact same aspects across the two studies (low aspect stability), they appeared to have sampled and listed aspects from a pool of idiosyncratic experiences with comparable strength of evidence (high evidence stability). Crucially, changes in the strength of evidence were systematically associated with changes in self-reported risk preferences. In sum, our analyses suggest that people's internal sampling process results in the retrieval of aspects that yield high evidence stability—thus providing a cognitive explanation for why self-reported risk preferences remain stable across time.

Cross-Study Analysis

Finally, to further clarify the predictive power of the strength and weight of evidence of the listed aspects, we repeated the analyses reported in study 1 by focusing on *cross-study pre- and post-dictions*. These analyses were particularly targeted at ruling out the possibility that the high predictive power of the aspects' strength and weight of evidence resulted from a methodological artifact; namely, that the respective ratings were provided in close proximity to the self-reported risk preferences. Thus, being able to predict (i.e., from study 1 to study 2) and post-dict (i.e., from study 2 to study 1) participants' self-reports of their risk preferences only using the aspects listed in the other study would constitute substantial evidence for the robustness of our main findings. Naturally, these tests rest on the assumption that people's risk preferences remain at least somewhat stable across time, a finding that has repeatedly been documented (e.g., Frey et al., 2017; Mata et al., 2018).

Table 3
Goodness of Fit Indicators of the Different Ordinal Regression Models in the Cross-Study Analyses.

Model	Accuracy	κ	Distinct Predictions
Fitting in study 1, prediction to study 2			
SoE	.28	.21	7
WoE	.29	.22	5
5 soc. dem. pred.	.26	.19	3
Sex	.21	.13	2
Age	.21	.13	3
Fitting in study 2, post-diction to study 1			
SoE	.35	.29	7
WoE	.31	.25	6
5 soc. dem. pred.	.24	.16	3
Sex	.22	.14	3
Age	.21	.14	1

Note: Predictions are based on ordinal regression models fit within study 1 and 2 (shown in Table 2). Accuracy = The proportion of correctly predicted categories (ratings between 0 and 10). κ = Cohen's kappa, with a chance level of 1/11. Distinct Predictions = The number of distinct/unique predictions made by a model (all numbers from 0 to 10 occurred in the empirical data). SoE = Mean strength of evidence per participant as predictor. WoE = Mean weight of evidence per participant as predictor. 5 soc. dem. pred. = Age, sex, years of education, income, and employment status as predictors. Sex = Sex as predictor. Age = Age as predictor.

Methods

Just as in study 2, we relied on the data of the 150 participants who completed both studies for this cross-study analysis. We again implemented the two-fold approach used in study 1; that is, we performed the cross-study analyses both with our initial set of six models, as well as with the ordinal regression models. To this end, we relied on the aspects (and estimated model parameters) obtained in study 1 (study 2) to generate predictions for the self-reported risk preference of study 2 (study 1).

Results

As can be seen in Figure 1, in the cross-study analyses the predictive accuracies of the six initial models were still substantial, with r_s ranging from .60 to .71. Specifically, the correlations between model predictions (from study 1) and self-reports (in study 2) were $r_s = .71$ (VUM), $r_s = .62$ (EXT), $r_s = .63$ (QT), $r_s = .65$ (FIRST), $r_s = .63$ (LAST), and $r_s = .64$ (SUM). Moreover, the correlations between model post-dictions (from study 2) and self-reports (in study 1) were $r_s = .70$ (VUM), $r_s = .61$ (EXT), $r_s = .60$ (QT), $r_s = .64$ (FIRST), $r_s = .60$ (LAST), and $r_s = .64$ (SUM). Hence, the predictive performance of all six models still substantially exceeded the predictive accuracy of the reference models.

Regarding the ordinal regression models, using the strength of evidence as predictor again led to the best model performance even in cross-study predictions, with correlations between model predictions (from study 1) and self-reports (in study 2) of $r_s = .72$ (SoE), and $r_s = .68$ (WoE), and correlations between model post-dictions (from study 2) and self-reports (in study 1) of $r_s = .68$ (SoE), and $r_s = .65$ (WoE). Moreover, also in terms of the accuracy, the model using the strength of evidence as predictor led to the best model performance when post-dicting from study 2 to study 1. However, when predicting from study 1 to study 2, the accuracies of the strength of evidence and the weight of evidence models were virtually identical (see Table 3). Finally, these two models clearly outperformed the reference models.

In the cross-study analyses, the proportion of identical predictions between these two models was slightly higher as compared to the within study analyses (see Table S3), yet still not at the upper bound (i.e., where each correct prediction of the worse model aligns with those of the better model) and thus still distinct in several cases (see also Figure S5). This is again highlighted in the larger number of distinct predictions made by the model with the strength of evidence as predictor, as opposed to the one with the weight of evidence as predictor (see Table 3).

Discussion

The cross-study analyses corroborated the conclusions drawn in study 1; namely, that the aspects' strength of evidence is the most important property of evidence for predicting self-reported risk preferences. As such, these analyses permitted ruling out a potential methodological confound due to the close temporal proximity between the section during which participants listed their aspects, and the section in which they self-reported their risk preferences.

Of note, although self-reported risk preferences showed a very high test–retest reliability across the two studies, some degree of intraindividual variability occurred. In light of this observation, some drop in model performance is naturally to be expected when making cross-study pre- and post-dictions. Notwithstanding this, and crucially, the models using the properties of evidence as predictors clearly outperformed the three reference models.

General Discussion

In the two studies presented in this article, we aimed to shed light on the information-integration processes underlying people's self-reports of their risk preferences, and to examine people's cognitive representations thereof. To this end, we made use of the process-tracing method of aspect listing and employed cognitive modeling to examine the extent to which different properties of evidence of the retrieved aspects are predictive of people's self-reports. Moreover, we investigated the stability of the "cognitive input" supposedly underlying people's self-reports (i.e., aspect- and evidence stability), the stability of the output (i.e., self-reported risk preferences), as well as the relation between stability in input and output. The results suggest three main take-home messages.

First, the two studies provide evidence for the internal validity of people's self-reports of their risk preferences. The desirable psychometric properties of the respective measures have increasingly been documented in recent research (e.g., Frey et al., 2017; Frey, Richter, et al., 2020; Mata et al., 2018), and the current analyses suggest a set of reasons for these observations. Specifically, people's self-reports appear to be the systematic result of a quantifiable information-integration process (see also Jarecki & Wilke, 2018): The aspects that participants retrieved from their memory during this process proved to be highly predictive for their self-reports—within and across the two studies reported here. Moreover, the aspects that form the input to this judgment-formation process mostly comprise situations that people frequently experience in their daily lives (see also Arslan et al., 2020; Schimmack et al., 2002; van der Linden, 2014; Weber, 2006)—rather than rare and exceptional, and thus potentially less diagnostic experiences.

Second, our model comparison unveiled several quantitative and qualitative properties of this information-integration process. From a theoretical point of view, people may consider three different properties of the retrieved information, namely, the weight, strength, and order of evidence. Whereas some research has primarily explored the weight of evidence of retrieved information (i.e., "how many pieces of information support a particular judgment?"; Jarecki & Wilke, 2018), here we also took into account the role of the other two dimensions. Our results indicated that the order of evidence may be largely irrelevant in this context, and people appeared to be particularly sensitive to the strength of evidence of retrieved information; that is, *how strongly* different aspects support a particular judgment concerning their risk preferences. This observation resonates with findings from other domains of judgment and decision making (Griffin & Tversky, 1992; Kvam & Pleskac, 2016) and suggests that similar information-integration processes may operate in judgment formation based on internal and external samples.

Third, our longitudinal analyses across the two studies illustrated that the properties of the cognitive input in people's judgments remained considerably stable (i.e., evidence stability), thus providing an explanation for why self-report measures of risk preference may show a high test–retest reliability (i.e., substantially higher than behavioral measures of the same construct; Frey et al., 2017; Lönnqvist et al., 2015; Mata et al., 2018). Specifically, the extent to which the strength of evidence of participants' listed aspects changed across time was strongly associated with changes in their self-reported risk preferences. The process of rendering self-reports arguably involves drawing internal samples of idiosyncratic experiences and past behaviors. According to our analyses, people retrieve aspects that are quite diverse in terms of their specific content, but highly similar in terms of their strength of evidence—and it was the latter dimension that people were mostly sensitive to when rendering a self-report. The high degree of evidence stability suggests that the retrieved experiences, albeit diverse, tend to support a similar degree of risk seeking or risk avoidance. In short, when rendering self-reports people may internally aggregate over different situations, and because the resulting self-reports thus encompass diverse settings, they may end up being predictive for a wide range of future behaviors and outcomes (e.g., Duckworth, Gendler, & Gross, 2016; Duckworth & Yeager, 2015). This interpretation

likely extends beyond self-reports of risk preference to domain-specific conceptions of risk preferences, and may also apply in other areas of psychological research (e.g., Blais & Weber, 2006; Duckworth & Kern, 2011; Eisenberg et al., 2019; Jarecki & Wilke, 2018; Sharma, Markon, & Clark, 2014; Wilke et al., 2014).

Cognitive Modeling as a Tool to Unpack Self-Reports?

As the three take-home messages above illustrate, we believe that our approach of using a process-tracing method—along with cognitive modeling—was highly instrumental in uncovering the information-integration processes and cognitive representations underlying people's self-reports. This approach rests on the assumption that judgment and decision making typically involve information-sampling and -integration processes, with information being sampled from either internal or external sources (Fiedler & Juslin, 2005; E. J. Johnson et al., 2007; Juslin & Olsson, 1997). Yet, to what extent can one be confident of having identified the true underlying process? Clearly, the various models implemented here remain approximations of the true psychological processes that may operate in people's minds, and even good model predictions do not guarantee that one has identified the "correct" process (see Roberts & Pashler, 2000). Hence, by increasing the degree of observable data beyond the self-reported aspects we used as input in our approach (e.g., reaction times, physiological indicators), lower-level and more fine-grained inferences concerning specific cognitive processes will become possible.

Nevertheless, we believe that the clear systematicity with which aspects and self-reported risk preference were related (within and across studies), the pattern with which stability in the aspects' strength of evidence was associated with stability in self-reported risk preference, and finally, the strong agreement in the strength of evidence as indicated by participants and by external raters are all indicators for the robustness of the approach implemented here. That said, in what follows we would like to discuss potential limitations of our studies and suggest avenues for further research in the future.

Limitations and Further Research

Aspect listing. One potential issue of aspect listing—at least when implemented in the traditional way (i.e., within one session only)—consists of the close temporal proximity between the listing of aspects and providing the self-report itself, hence potentially inflating the respective consistency. Our design with a retest study permitted addressing this issue directly: Even in the cross-study analyses the predictive accuracies of the various models were high and far superior compared to those of sociodemographic predictors. This suggests that the good performance of the cognitive models does not merely reflect a methodological artifact.

Yet, there are potentially even more fundamental issues related to the method of aspect listing that are worthy of a careful discussion. As outlined in our introduction, a basic motivation for employing aspect listing is to avoid having to prompt respondents to engage in introspection *in hindsight*; that is, to reflect on how they had rendered a previous self-report. Specifically, it has been argued that such retrospective

metacognitive judgments may be unreliable, as people lack sufficient insight into the cognitive processes underlying their own judgments (Nisbett & Bellows, 1977; Nisbett & Wilson, 1977). Thus, to avoid this potential issue, methods such as aspect listing or think-aloud protocols aim to trace information processing on the fly (e.g., Ericsson & Simon, 1980, 1993). Naturally, there are also some intricacies with this approach, as it evidently rests on the assumption that people are capable of providing veridical reports of their own, ongoing thoughts—and this assumption may not always be met, at least not entirely: On the one hand, the task of sequentially typing in one's ongoing thoughts may alter the judgment-formation process to be more systematic, thus potentially leading to a more structured way of rendering a self-report (see Ericsson & Simon, 1980; Fox, Ericsson, & Best, 2011). To illustrate, the somewhat stronger bimodal distribution of participants' self-reports in our studies (i.e., as compared to in previous studies; e.g., Dohmen et al., 2011; Frey et al., 2017) might be a manifestation of this possibility—although there were no indications for systematic mean differences, depending on whether self-reports were provided after or before the actual aspect listing (as investigated in a pilot study, see methods section of study 1). On the other hand, assuming that the method of aspect listing does not overly distort the ongoing judgment-formation process, one still cannot be entirely sure that the listed aspects reflect fully accurate memories, as memories of everyday life events could be altered and transformed (for reviews, see Koriat, 2007; Koriat et al., 2000). Thus, in future research it will be useful to test whether our findings also hold for other process-tracing methods such as think-aloud protocols, which might be more robust in this regard (Fox et al., 2011). Relatedly, it may be worthwhile to test the extent to which particular contexts trigger the retrieval of specific (classes of) aspects, which could in principle explain why aspect stability (but not evidence stability) was low across the two studies conducted here. Taken together, people may not always have direct introspective access to the processes involved in their judgments and decisions (i.e., particularly when being prompted to reflect on such processes explicitly and in hindsight). Yet, under certain conditions and when using the appropriate methods they may indeed be able to report on their current thoughts quite accurately, thus providing reliable insight concerning the underlying cognitive processes (e.g., Adair & Spinner, 1981; Berger, Dennehy, Bargh, & Morsella, 2016; Ericsson & Simon, 1980, 1993; Hurlburt & Heavey, 2001; White, 1980).

Modeling approach. We have sampled diverse models from the literature on judgment and decision making that describe manifold information-integration processes, and which cover a wide space ranging from simple heuristics to learning models. As the empirical data imposed some constraints concerning the level of detail with which fine-grained model comparisons were possible, we additionally relied on a more general model comparison—focusing on the distinction between the strength of evidence and the weight of evidence. In the future, the employment of yet other process-tracing approaches (see points discussed above) might allow for more fine-grained analyses in this respect.

Moreover, as we modeled one self-report per participant, we estimated the free parameters across individuals. Although this is a widespread procedure in various applications of cognitive modeling (e.g., Birnbaum, 2008; Erev, Ert, Plonsky, Cohen,

& Cohen, 2017; Erev et al., 2010), this approach is not without its problems. For example, not all participants may rely on the same information-integration processes (e.g., Frey, Rieskamp, & Hertwig, 2015; Mata, von Helversen, & Rieskamp, 2010; Payne et al., 1988) and/or may be best described with the same parameter values (e.g., Kellen et al., 2016; Pedroni et al., 2017). In short, it is unclear to what extent findings based on the *interindividual* level generalize to the *intraindividual* level (Molenaar, 2004; Molenaar & Campbell, 2009), and future research is thus needed to clarify a potential heterogeneity between different persons' cognitive processes.

Outcome measures. Finally, future research may also investigate to what extent our findings extend to self-reports of domain-specific risk preferences. Jarecki and Wilke (2018) have examined how cognitive processes potentially vary across different (evolutionary) content-domains (see also Wilke et al., 2014). Similar analyses, yet including models that take into account the strength of evidence of retrieved information, could thus also be conducted for domain-specific risk preferences as are often assessed in psychological research (Rolison & Shenton, 2020; Weber et al., 2002). One may expect that aspects retrieved for specific domains of life (e.g., recreation, health, finance) may be more heterogeneous across domains, but more homogeneous within, as compared to those retrieved in response to a domain-general question as investigated here—which may ultimately increase aspect stability.

Conclusions

Zooming out, our approach to modeling people's self-reported risk preferences involves several contributions that inform psychological assessment in general, and provides theoretical and measurement-related insight into the construct of risk preference more specifically.

First, we bridged two methodological approaches that are too often employed separately; that is, we investigated self-reported preferences (as typically employed in psychometric research relying on questionnaires) by implementing cognitive modeling using a range of different models. Integrating these approaches proved helpful for a better understanding of the construct validity of self-reported risk preference, and we hope that our approach will inspire similar applications in other areas of psychological research in the future.

Second, our investigations provide substantial evidence that self-reports of risk preference are robustly rooted in people's idiosyncratic experiences, and are thus internally valid. Specifically, the desirable psychometric properties of respective self-report measures—here tapping risk preference, but potentially also in the case of self-reports of other psychological constructs—may emerge as the result of an information-integration process that aggregates multiple samples that people draw from their autobiographical memory.

Third and finally, our findings have an important implication for applied settings: Risk preferences can have a dramatic impact on important life outcomes, and are thus frequently assessed in various real-life contexts, such as concerning health- and safety-related matters or when designing investment portfolios. In doing so, the tool of choice is often one of numerous self-report measures. These measures may be

not only frugal in their application—but according to our findings also sound from a psychological perspective.

References

Adair, J. G., & Spinner, B. (1981). Subjects' Access to Cognitive Processes: Demand Characteristics and Verbal Report. *Journal for the Theory of Social Behaviour*, *11*(1), 31–52. doi: 10.1111/j.1468-5914.1981.tb00021.x

Andreoni, J., & Kuhn, M. A. (2019). Is it safe to measure risk preferences? Assessing the completeness, predictive validity, and measurement error of various techniques. *Working Paper*. Retrieved from `https://static1.squarespace.com/static/5c79b3d29b8fe82f5cb96360/t/5cc0debb71c10bd5d9ab45f3/1556143804348/mCRB_WP.pdf`

Appelt, K. C., Hardisty, D. J., & Weber, E. U. (2011). Asymmetric discounting of gains and losses: A query theory account. *Journal of Risk and Uncertainty*, *43*(2), 107–126. doi: 10.1007/s11166-011-9125-1

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. doi: 10.1038/s41598-020-72077-5

Balatel, A., Boero, R., Jonaityte, I., Monti, M., Novarese, M., & Pacelli, V. (2013). Beyond the MiFID: Envisioning cognitively suitable and representationally supportive approaches to assessing investment preferences for more informed financial decisions. *CAREFIN Occasional Paper*. Retrieved from `http://hdl.handle.net/11858/00-001M-0000-0024-ED29-9`

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. doi: 10.1016/j.jml.2012.11.001

Barron, G., & Yechiam, E. (2009). The coexistence of overestimation and underweighting of rare events and the contingent recency effect. *Judgment and Decision Making*, *4*, 447–460. Retrieved from `http://journal.sjdm.org/9729b/jdm9729b.pdf`

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. doi: 10.1007/s11166-017-9261-3

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*(5), 961–981. doi: 10.1287/opre.30.5.961

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200. doi: 10.1037/h0024835

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. doi: 10.1073/pnas.0500333102

Berger, C. C., Dennehy, T. C., Bargh, J. A., & Morsella, E. (2016). Nisbett and Wilson (1977) Revisited: The Little That We Can Know and Can Tell. *Social Cognition*, *34*(3), 167–195. doi: 10.1521/soco.2016.34.3.167

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463 - 501. doi: 10.1037/0033-295X.115.2.463

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

Broomell, S. B., Budescu, D. V., & Por, H.-H. (2011). Pair-wise comparisons of multiple models. *Judgment and Decision Making*, *6*(8), 821–831. Retrieved from `http://journal.sjdm.org/11/m09/m09.html`

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. doi: 10.1177/1745691610393980

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*(3), 367–383. doi: 10.1037/0033-2909.125.3.367

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via amazon mturk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*, 2156 - 2160. doi: 10.1016/j.chb.2013.05.009

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475-494. doi: 10.1177/001316444600600405

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. doi: 10.1037/0003-066X.34.7.571

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95–106. doi: 10.1037/h0037613

Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, *95*(3), 542-575. doi: 10.1037/0033-2909.95.3.542

Dohmen, T. J., 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*, 522–550. doi: 10.1111/j.1542-4774.2011.01015.x

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational strategies for self-control. *Perspectives on Psychological Science*, *11*(1), 35-55. doi: 10.1177/1745691615623247

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*, 259-268. doi: 10.1016/j.jrp.2011.02.004

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087-1101. doi: 10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237-251. doi: 10.3102/0013189X15584327

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., 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*, 2319. doi: 10.1038/s41467-019-10301-1

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*(4), 369–409. doi: 10.1037/rev0000062

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., ... Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15–47. doi: 10.1002/bdm.683

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*,

87(3), 215–251. doi: 10.1037/0033-295X.87.3.215

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. doi: 10.3758/BRM.41.4.1149

Ferrarini, G., & Wymeersch, E. (2006). *Investor protection in Europe: Corporate law making, the MiFID and beyond*. Oxford, UK: Oxford University Press.

Fiedler, K., & Juslin, P. (2005). *Information sampling and adaptive cognition*. New York: Cambridge University Press.

Fiedler, K., Renn, S.-Y., & Kareev, Y. (2010). Mood and judgments based on sequential sampling. *Journal of Behavioral Decision Making*, *23*(5), 483–495. doi: 10.1002/bdm.669

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. doi: 10.1007/BF00143739

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. doi: 10.1037/a0021663

Frey, R., Duncan, S., & Weber, E. U. (2020). Towards a typology of risk preference: Four risk profiles describe two thirds of individuals in a large sample of the U.S. population. *PsyArXiv Preprint*. doi: 10.31234/osf.io/yjwr9

Frey, R., Mata, R., & Hertwig, R. (2015). The role of cognitive abilities in decisions from experience: Age differences emerge as a function of choice set size. *Cognition*, *142*, 60–80. doi: 10.1016/j.cognition.2015.05.004

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*, e1701381. doi: 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*. doi: 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. doi: 10.1037/a0038118

Galizzi, M. M., Machado, S. R., & Miniaci, R. (2016). Temporal stability, cross-validity, and external validity of risk preferences measures: Experimental evidence from a UK representative sample. *London School for Economics and Political Science Working Paper*. doi: 10.2139/ssrn.2822613

Galton, F. (1874). *English men of science*. London: Macmillan.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. doi: 10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. doi: 10.1037/0033-295X.103.4.650

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In G. Gigerenzer, P. M. Todd, & The ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 75–95). New York: Oxford University Press.

Gluth, S., & Jarecki, J. B. (2019). On the Importance of Power Analyses for Cognitive Modeling. *Computational Brain & Behavior*, *2*(3-4), 266–270. doi: 10.1007/s42113-019-00039-w

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). *rstanarm: Bayesian applied regression modeling via Stan.* Retrieved from `http://mc-stan.org/` (R package version 2.17.4)

Greifeneder, R., & Schwarz, N. (2014). Metacognitive processes and subjective experiences. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 314–327). New York: The Guilford Press.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435. doi: 10.1016/0010-0285(92)90013-R

Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review*, *9*(2). doi: 10.2307/2086306

Hammond, K. R., & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications.* New York: Oxford University Press.

Hardisty, D. J., Johnson, E. J., & Weber, E. U. (2010). A dirty word or a dirty world? attribute framing, political affiliation, and query theory. *Psychological Science*, *21*, 86–92. doi: 10.1177/0956797609355572

Hastie, R., & Dawes, R. M. (2001). A general framework for judgment. In *Rational choice in an uncertain world: The psychology of judgment and decision making* (pp. 47–69). Thousand Oaks, CA: Sage Publications, Inc.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review*, *93*(3), 258–268. doi: 10.1037/0033-295X.93.3.258

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*, 534–539. doi: 10.1111/j.0956-7976.2004.00715.x

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The role of information sampling in risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 72–91). New York: Cambridge University Press.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237. doi: 10.1016/j.cognition.2009.12.009

Highhouse, S., & Gallo, A. (1997). Order effects in personnel decision making. *Human Performance*, *10*, 31-46. doi: 10.1207/s15327043hup1001_2

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1-55. doi: 10.1016/0010-0285(92)90002-J

Hurlburt, R. T., & Heavey, C. L. (2001). Telling what we know: Describing inner experience. *Trends in Cognitive Sciences*, *5*(9), 400–403. doi: 10.1016/S1364-6613(00)01724-1

Jarecki, J. B., & Wilke, A. (2018). Into the black box: Tracing information about risks related to 10 evolutionary problems. *Evolutionary Behavioral Sciences*, *12*, 230–244. doi: 10.1037/ebs0000123

Jobe, J. B. (2000). Cognitive processes in self-report. In A. A. Stone, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Impli-*

cations for research and practice (pp. 25–29). Mahwah, NJ: Lawrence Erlbaum.

Jobe, J. B. (2003). Cognitive psychology and self-reports: Models and methods. *Quality of Life Research*, *12*, 219–227. doi: 10.1023/A:1023279029852

Jobe, J. B., & Herrmann, D. J. (1996). Implications of models of survey cognition for memory theory. In D. J. Herrman, C. McEvoy, C. Hertzog, P. Hertel, & M. K. Johnson (Eds.), *Basic and applied memory research* (pp. 193–205). New York: Laurence Erlbaum Associates, Inc.

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 461–474. doi: 10.1037/0278-7393.33.3.461

Johnson, J. G., & Raab, M. (2003). Take the first: Option-generation and resulting choices. *Organizational Behavior and Human Decision Processes*, *91*, 215–229. doi: 10.1016/S0749-5978(03)00027-X

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*, 344-366. doi: 10.1037/0033-295X.104.2.344

Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences*, *107*(38), 16489–16493. doi: 10.1073/pnas.1011492107

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. doi: 10.2307/1914185

Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in)variant are subjective representations of described and experienced risk and rewards? *Cognition*, *157*, 126–138. doi: 10.1016/j.cognition.2016.08.020

Kendall, M. G. (1948). *Rank correlation methods*. London, UK: Griffin.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609–639. doi: 10.1037/0033-295X.100.4.609

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The cambridge handbook of consciousness* (pp. 289–325). New York: Cambridge University Press.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology*, *51*(1), 481–537. doi: 10.1146/annurev.psych.51.1.481

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition*, *152*, 170–180. doi: 10.1016/j.cognition.2016.04.008

Lejarraga, T., Frey, R., Schnitzlein, D. D., & Hertwig, R. (2019). No effect of birth order on adult risk taking. *Proceedings of the National Academy of Sciences*, *116*, 6019-6024. doi: 10.1073/pnas.1814153116

Lerner, J. S., Gonzalez, R. M., Small, D. A., & Fischhoff, B. (2003). Effects of fear and anger on perceived risks of terrorism: A national field experiment. *Psychological Science*, *14*(2), 144–150. doi: 10.1111/1467-9280.01433

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*, 55-55.

Lindskog, M., Winman, A., & Juslin, P. (2013). Naïve point estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 782-800. doi: 10.1037/a0029670

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267-286. doi: 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. doi: 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. doi: 10.2307/2232669

Lopes, L. L. (1984). Risk and distributional inequality. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 465-485. doi: 10.1037/0096-1523.10.4.465

Maccrimmon, K. R., & Wehrung, D. A. (1985). A portfolio of risk measures. *Theory and Decision*, *19*(1), 1–29. doi: 10.1007/BF00134352

Mamerow, L., Frey, R., & Mata, R. (2016). Risk taking across the life span: A comparison of self-report and behavioral measures of risk taking. *Psychology and Aging*, *31*, 711–723. doi: 10.1037/pag0000124

Mata, R., Frey, R., Richter, D., Schupp, J., & Hertwig, R. (2018). Risk preference: A view from psychology. *Journal of Economic Perspectives*, *32*(2), 155–172. doi: 10.1257/jep.32.2.155

Mata, R., Josef, A. K., & Hertwig, R. (2016). Propensity for risk taking across the life span and around the globe. *Psychological Science*, *27*(2), 231–243. doi: 10.1177/0956797615617811

Mata, R., von Helversen, B., & Rieskamp, J. (2010). Learning to choose: Cognitive aging and strategy selection learning in decision making. *Psychology and Aging*, *25*(2), 299. doi: 10.1037/a0018923

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81-90. doi: 10.1037/0022-3514.52.1.81

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*(6), 423–429. doi: 10.1111/j.1467-9280.1997.tb00455.x

Mishra, S., Barclay, P., & Sparks, A. (2017). The relative state model: Integrating need-based and ability-based pathways to risk-taking. *Personality and Social Psychology Review*, *21*(2), 176-198. doi: 10.1177/1088868316644094

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. doi: 10.1207/s15366359mea0204_1

Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*(2), 112–117. doi: 10.1111/j.1467-8721.2009.01619.x

Nicholson, N., Soane, E., Fenton-O'Creevy, M., & Willman, P. (2005). Personality and domain-specific risk taking. *Journal of Risk Research*, *8*(2), 157–176. doi: 10.1080/1366987032000123856

Nisbett, R. E., & Bellows, N. (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. *Journal of Personality*, *35*(9), 613–624. doi: 10.1037/0022-3514.35.9.613

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. doi: 10.1037/0033-295X.84.3.231

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. doi: 10/10630a/

jdm10630a

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 534–552. doi: 10.1037/0278-7393.14.3.534

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, *1*, 803–809. doi: 10.1038/s41562-017-0219-x

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior Research Methods*, *46*(4), 1023–1031. doi: 10.3758/s13428-013-0434-y

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367. doi: 10.1037/0033-295X.107.2.358

Rohrer, J. M. (2017). Test–retest reliabilities of scales included in the socio-economic panel study. *PsychArxiv Preprint*. doi: 10.31219/osf.io/3ncbt

Rolison, J. J., & Shenton, J. (2020). How much risk can you stomach? Individual differences in the tolerance of perceived risk across gender and risk domain. *Journal of Behavioral Decision Making*, *33*, 63–85. doi: 10.1002/bdm.2144

Schimmack, U., Diener, E., & Oishi, S. (2002). Life-satisfaction is a momentary judgment and a stable personality characteristic: The use of chronically accessible and stable sources. *Journal of Personality*, *70*(3), 345-384. doi: 10.1111/1467-6494.05008

Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., & Willemsen, M. C. (2017). Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science*, *26*(5), 442-450. doi: 10.1177/0963721417708229

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *The American Journal of Evaluation*, *22*(2), 127-160. doi: 10.1016/S1098-2140(01)00133-3

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. doi: 10.1037/a0034418

Sitkin, S. B., & Pablo, A. L. (1992). Reconceptualizing the determinants of risk behavior. *Academy of Management Review*, *17*, 9–38. doi: 10.5465/amr.1992.4279564

Steiner, M. D., & Frey, R. (in press). Representative design in psychological assessment: A case study using the balloon analogue risk task (BART). *Journal of Experimental Psychology: General*. Retrieved from `https://psyarxiv.com/dg4ks/`

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT press.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-286. doi: 10.1037/h0070288

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529-554. doi: 10.1086/214483

Tisdall, L., Frey, R., Horn, A., Ostwald, D., Horvath, L., Blankenburg, F., ... Mata, R. (2020). Brain-behavior associations for risk taking depend on the measures used to capture individual differences. *Frontiers in Behavioral Neuroscience*, *14*. doi: 10.3389/fnbeh.2020.587152

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*(5811), 515–518. doi: 10.1126/science.1134239

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

van der Linden, S. (2014). On the relationship between personal experience, affect and risk perception: The case of climate change. *European Journal of Social Psychology*, *44*, 430–440. doi: 10.1002/ejsp.2008

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Wang, X. T., Kruger, D. J., & Wilke, A. (2009). Life history variables and risk-taking propensity. *Evolution and Human Behavior*, *30*(2), 77–84. doi: 10.1016/j.evolhumbehav.2008.09.006

Weber, E. U. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change*, *77*, 103–120. doi: 10.1007/s10584-006-9060-3

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. doi: 10.1002/bdm.414

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A query-theory account. *Psychological Science*, *18*, 516–523. doi: 10.1111/j.1467-9280.2007.01932.x

Weber, E. U., & Milliman, R. A. (1997). Perceived risk attitudes: Relating risk perception to risky choice. *Management Science*, *43*(2), 123–144. doi: 10.1287/mnsc.43.2.123

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review*, *87*(1), 105–112.

Wilke, A., Sherman, A., Curdt, B., Mondal, S., Fitzgerald, C., & Kruger, D. J. (2014). An evolutionary domain-specific risk scale. *Evolutionary Behavioral Sciences*, *8*(3), 123–141. doi: 10.1037/ebs0000011

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. doi: 10.1177/1745691617693393

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review*, *12*, 387–402. doi: 10.3758/BF03193783

Zuckerman, M. (2002). *Sensation Seeking and Risky Behavior*. Binghamton, NY: Maple-Vail Press.

Appendix B: Steiner & Frey (in press)

Steiner, M. D., & Frey, R. (in press). Representative design in psychological assessment: A case study using the balloon analogue risk task (BART). *Journal of Experimental Psychology: General*. Retrieved from `https://psyarxiv.com/dg4ks/`

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

Markus D. Steiner, Renato Frey
University of Basel

Abstract

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 adaptions. 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

Various psychological assessments are routinely performed by means of behavioral tasks, including the measurement and modeling of individual differences in risk taking (Frey, Pedroni, Mata, Rieskamp, & Hertwig, 2017; Frey, Richter, Schupp, Hertwig, & Mata, 2020; Lauriola, Panno, Levin, & Lejuez, 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, Choi, Laibson, & Madrian, 2008; Charness, Gneezy, & Imas, 2013), recent evidence has highlighted substantial psychometric limitations of this measurement approach (e.g., Beauchamp, Cesarini, & Johannesson, 2017; Berg, Dickhaut, & McCabe, 2005; Eisenberg et al., 2019; Frey et al., 2017; Lönnqvist, Verkasalo, Walkowitz, & Wichardt, 2015; Millroth, Juslin, Winman, Nilsson, & Lindskog, 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, Seitz, & Frey, 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, Korczykowski, Pluta, Hoang, & Detre, 2008; Schonberg, Fox, & Poldrack, 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, Markon, & Clark, 2014; Wallsten, Pleskac, & Lejuez, 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, Bardy, Smart, & Pagulayan, 2003; for an overview see Araújo, Davids, & Passos, 2007 and Dhami, Hertwig, & Hoffrage, 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 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, Frey, Richter, Schupp, & Hertwig, 2018; Slovic, 1962)—and thus ultimately hamper a successful assessment of meaningful individual differences. This article illustrates this argument, and systematically

---

Markus D. Steiner, Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel; Renato Frey, Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel.

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. 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. Online supplemental materials, 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.

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, Daughters, Hernandez, Richards, & Lejuez, 2005; Hopko et al., 2006; Hunt, Hopko, Bare, Lejuez, & Robinson, 2005), as well as in applied contexts (e.g., Aklin, Lejuez, Zvolensky, Kahler, & Gwadz, 2005; Lejuez, Aklin, Zvolensky, & Pedulla, 2003). For instance, the BART has been used to predict interindividual differences in substance use (e.g., Campbell, Samartgis, & Crowe, 2013; Hanson, Thayer, & Tapert, 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, Hau, Papassotiropoulos, & Hertwig, 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 decision 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, Rieskamp, & Hertwig, 2015; Hertwig, Barron, Weber, & Erev, 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, Conradt, Leuker, & Hertwig, 2020; Pleskac & Hertwig, 2014).

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 et al., 2007; Lejuez, Aklin, Jones, et al., 2003; Skeel, Pilarski, Pytlak, & Neudecker, 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, Frey, & Pleskac, 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, Lejuez, & de Wit, 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, Wallsten, Wang, & Lejuez, 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 behaviors (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)

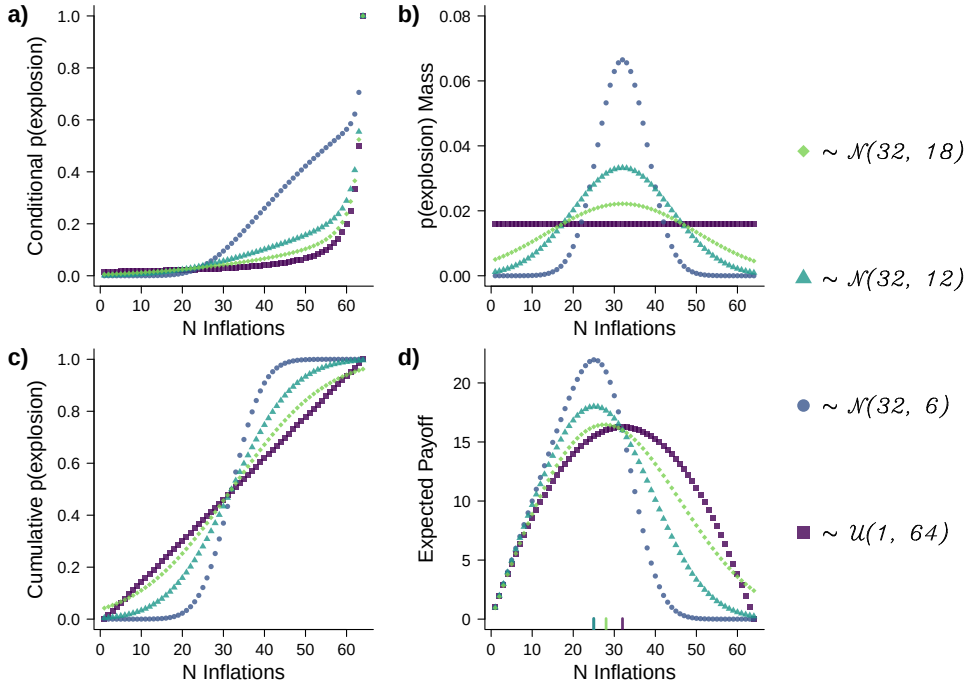

Figure 1. Illustration of four different task designs, each implementing a different stochastic structure in the BART. Colors/shapes indicate different distributions of explosion points, with purple/squared dots depicting the standard implementation of the BART (i.e., uniform distribution) and the other 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 $\text{BART}_{\text{normal-L}}$.

is defined as

$$p(expl_i|\neg expl_{i-1}) = 1/(C - i + 1) \tag{1}$$

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., $p(expl_i) = 1/C$ for all inflation stages $i \in \{1, 2, \ldots, C\}$).

Evidently, the typical implementation of the BART—hereinafter referred to as the $\text{BART}_{\text{uniform}}$—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

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

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?

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 (Dhami 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 $\text{BART}_{\text{uniform}}$. 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 $\text{BART}_{\text{uniform}}$, 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.

Second, over-learning one's prior expectations may be especially challenging in the case of the $\text{BART}_{\text{uniform}}$ 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, Wright, & Rakow, 2014). Thus, this second issue likely aggravates the consequences of the first issue.

Third, the payoff-maximizing behavior in the $\text{BART}_{\text{uniform}}$ consists of inflating

[2]Note 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.

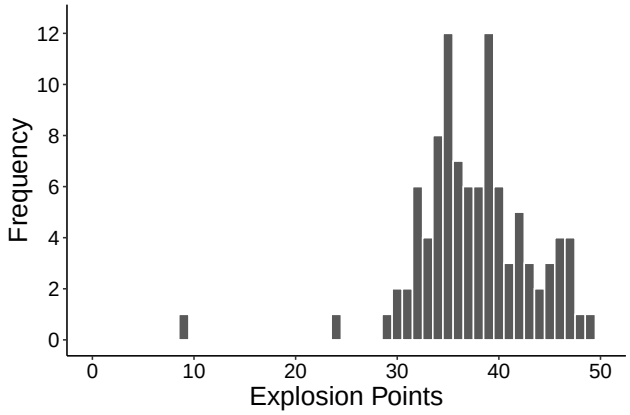

Figure 2. Distributions of explosion points of 100 real balloons, inflated with a bicycle pump.

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). Consequently, 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.

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 design, 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).

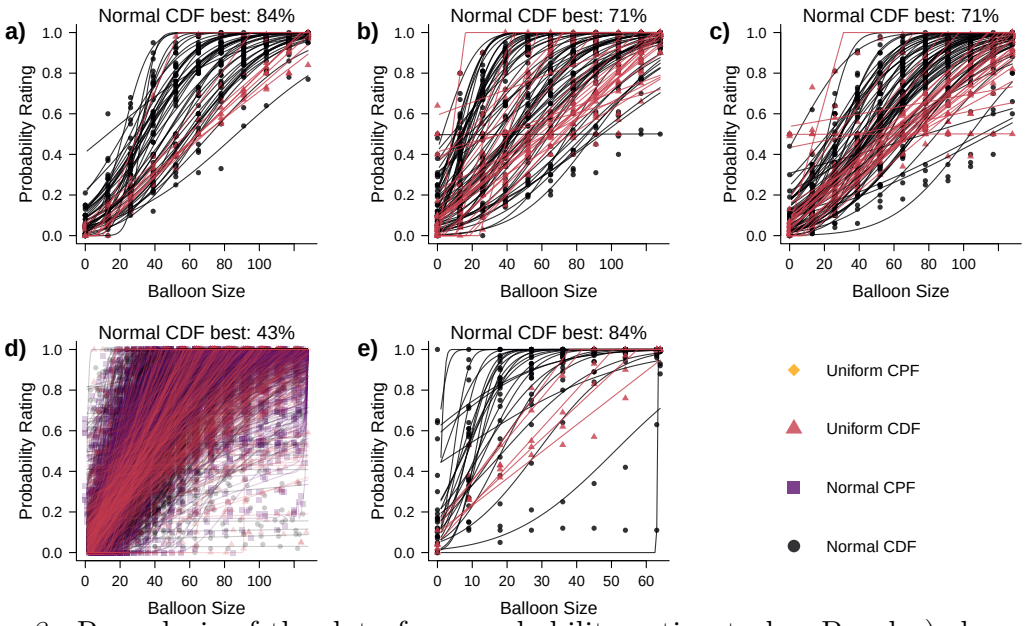

Figure 3. Reanalysis of the data from probability rating tasks. Panel a) shows the result of the reanalysis of study 1 from Schürmann et al. (2018). Panels b) and c) show the results of the reanalysis of study 2 from Schürmann et al. (2018) after participants have played one trial, and all 30 trials, 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.

Results

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

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

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 $\text{BART}_{\text{uniform}}$: Most participants expected a normal distribution 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., $\text{BART}_{\text{normal}}$). We expected this change in the task architecture to lead to several improvements:

First, because explosion points are clustered in the $\text{BART}_{\text{normal}}$, 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 $\text{BART}_{\text{uniform}}$ 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 $\text{BART}_{\text{uniform}}$, 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 $\text{BART}_{\text{uniform}}$.

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, Highhouse, & Nye, 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 $\text{BART}_{\text{uniform}}$ (N = 190), $\text{BART}_{\text{normal-H}}$ (N = 195),

[4]We 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.

$\text{BART}_{\text{normal-M}}$ (N = 197), and $\text{BART}_{\text{normal-L}}$ (N = 190). In the $\text{BART}_{\text{uniform}}$ the balloons' explosion points were drawn from $\mathcal{U}(1, 64)$. In the three versions of the $\text{BART}_{\text{normal}}$, 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 ($\text{BART}_{\text{normal-H}}$), $\mathcal{N}(32, 18)$ representing a medium SD ($\text{BART}_{\text{normal-M}}$), and $\mathcal{N}(32, 6)$ representing a low SD ($\text{BART}_{\text{normal-L}}$). 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 $\text{BART}_{\text{uniform}}$ 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 $\text{BART}_{\text{normal}}$ (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, Manske, Preckel, & Wilhelm, 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 different 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, Blais, & Betz, 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 explosions 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, Gabry, Ali, & Brilleman, 2018). The contrasts were $\text{BART}_{\text{uniform}}$ vs. the three implementations 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.

[5] We 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.

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 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.

Priors and ROPEs. 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 $\mathcal{N}(mean(y), sd(y) * 1000)$ and $\mathcal{U}(sd(y)/1000, sd(y) * 1000)$ for μ and σ, and $\mathcal{E}(1/29)$ for ν, with $\nu \geq 1$. Finally, for correlations we used the prior $beta(3, 3)$.

As suggested by Makowski, Ben-Shachar, and Lüdecke (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 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 explosions 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 $\text{BART}_{\text{normal-L}}$ did most participants (64.61%) believe that the explosion points were normally distributed,

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

Comparison	Δ [95% HDI]	pROPE	*d*
Mean beliefs about the optimal behavior			
$BART_{normal\text{-}H}$ - $BART_{uniform}$	1.56 [-0.85, 4.07]	.036	0.16
$BART_{normal\text{-}M}$ - $BART_{uniform}$	0.27 [-1.95, 2.39]	.077	0.11
$BART_{normal\text{-}L}$ - $BART_{uniform}$	**3.61** [1.54, 5.64]	< .000	0.38
SD beliefs about the optimal behavior			
$BART_{normal\text{-}H}$ - $BART_{uniform}$	-1.81 [-4.51, 0.80]	.033	-
$BART_{normal\text{-}M}$ - $BART_{uniform}$	**-6.03** [-8.30, -3.84]	< .000	-
$BART_{normal\text{-}L}$ - $BART_{uniform}$	**-4.69** [-7.05, -2.38]	< .000	-
Deviance of participants' beliefs about optimal behavior from objectively optimal behavior			
$BART_{normal\text{-}H}$ - $BART_{uniform}$	**5.55** [3.09, 8.00]	< .000	0.47
$BART_{normal\text{-}M}$ - $BART_{uniform}$	**7.27** [5.03, 9.37]	< .000	0.69
$BART_{normal\text{-}L}$ - $BART_{uniform}$	**10.60** [8.52, 12.62]	< .000	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.

with an average rating of 29.01. In the other implementations, only the minority of 39.69% ($BART_{normal\text{-}M}$), 32.26% ($BART_{normal\text{-}H}$), 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 $BART_{normal\text{-}M}$, 19.15 in the $BART_{normal\text{-}H}$, 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 $BART_{normal\text{-}M}$ and the $BART_{normal\text{-}H}$ ($b = 1.70$, 95% HDI: [-0.82, 4.22,], pROPE = .382, $d = 0.17$), we found conclusive evidence that participants' ratings in the $BART_{normal\text{-}L}$ were higher than in the $BART_{normal\text{-}M}$ ($b = 8.16$, 95% HDI: [5.50, 10.61], pROPE < .000, $d = 0.63$). 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 $BART_{normal\text{-}L}$ 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 $\text{BART}_{\text{normal-M}}$ or in the $\text{BART}_{\text{normal-H}}$. 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 $\text{BART}_{\text{normal-L}}$ and in the $\text{BART}_{\text{normal-M}}$ had a smaller variability (i.e., across participants), as compared to the beliefs of participants in the $\text{BART}_{\text{uniform}}$ (see Table 1). Yet, there was no conclusive evidence whether or not participants in the $\text{BART}_{\text{normal-H}}$ and in the $\text{BART}_{\text{uniform}}$ 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 $\text{BART}_{\text{uniform}}$ than in the three implementations of the $\text{BART}_{\text{normal}}$ (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 $\text{BART}_{\text{normal}}$ as opposed to in the $\text{BART}_{\text{uniform}}$. Moreover, we predicted that the adjusted BART scores would be higher and exhibit less variability across participants within the $\text{BART}_{\text{normal}}$, the lower the standard deviation of the explosion points. In line with this prediction, we found conclusive evidence that, compared to the $\text{BART}_{\text{uniform}}$, the adjusted BART scores were higher in all three implementations of the $\text{BART}_{\text{normal}}$ (see Table 2; see Table S1 for the estimates of participants' adjusted BART scores). Within the $\text{BART}_{\text{normal}}$ and further in line with Hypothesis 3, there was conclusive evidence that the adjusted BART scores were higher in the $\text{BART}_{\text{normal-L}}$ than in the $\text{BART}_{\text{normal-M}}$. Yet, there was conclusive evidence that the adjusted BART scores in the $\text{BART}_{\text{normal-M}}$ were lower as compared to those in the $\text{BART}_{\text{normal-H}}$.

Also in line with Hypothesis 3, there was conclusive evidence that the adjusted BART scores exhibited less variability between participants in the $\text{BART}_{\text{normal-M}}$ and in the $\text{BART}_{\text{normal-L}}$, as compared to in the $\text{BART}_{\text{uniform}}$ (see Table 2). However, there was no conclusive evidence for whether the variability between the $\text{BART}_{\text{uniform}}$ and the $\text{BART}_{\text{normal-H}}$ differed. Yet, also in line with Hypothesis 3, the variability of the adjusted BART scores was lower in the $\text{BART}_{\text{normal-M}}$ than in the $\text{BART}_{\text{normal-H}}$, and lower in the $\text{BART}_{\text{normal-L}}$ than in the $\text{BART}_{\text{normal-M}}$.

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 $\text{BART}_{\text{uniform}}$ as compared to the three implementations of the

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

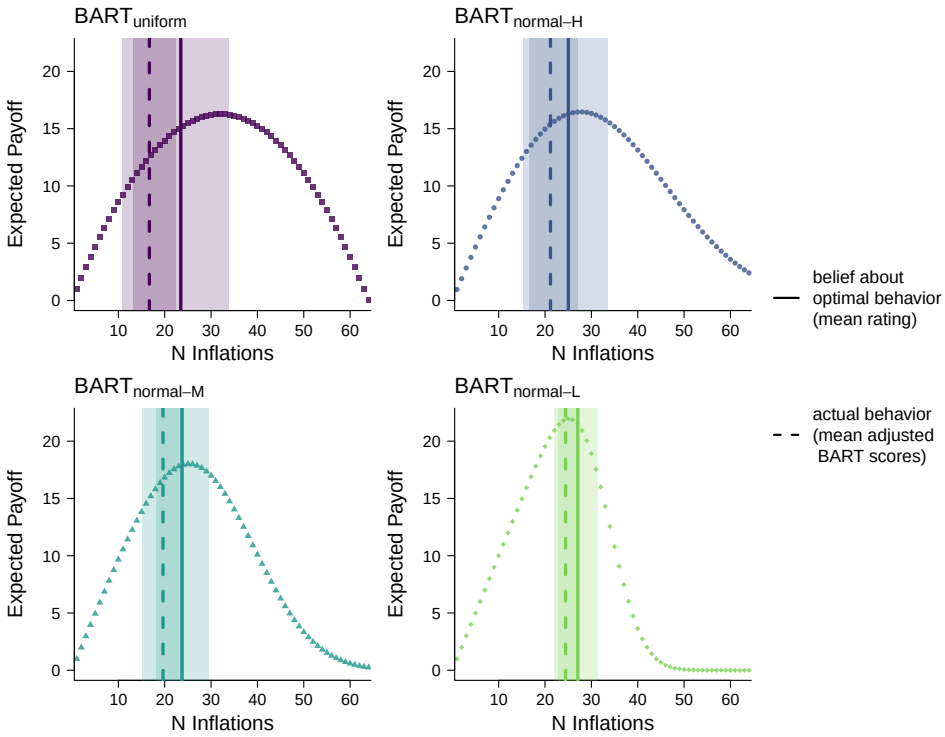

Figure 4. Participants' beliefs about the optimal behavior and their actual behavior. Vertical lines indicate the means across participants concerning their beliefs about the optimal behavior (solid lines) and their actual behavior (dashed lines). Shaded areas around the vertical lines indicate the standard deviations of participants' beliefs about the optimal behavior and of their adjusted BART scores.

$\text{BART}_{\text{normal}}$ (see Table 2). Moreover, within the $\text{BART}_{\text{normal}}$ the deviance between the adjusted BART scores and the objectively optimal behavior was larger in the $\text{BART}_{\text{normal-H}}$ as compared to the in the $\text{BART}_{\text{normal-M}}$ and larger in the $\text{BART}_{\text{normal-M}}$ as compared to in the $\text{BART}_{\text{normal-L}}$.

Convergent validity. In Hypothesis 4, we predicted that the $\text{BART}_{\text{normal}}$ would have a higher convergent validity as opposed to the $\text{BART}_{\text{uniform}}$. 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 $r = .08$ ($\text{BART}_{\text{uniform}}$), $r =$ -.05 ($\text{BART}_{\text{normal-H}}$), $r = .12$ ($\text{BART}_{\text{normal-M}}$), and $r = .04$ ($\text{BART}_{\text{normal-L}}$). The total number of explosions per participant was somewhat more strongly but still weakly related to the other measures, with average correlations of $r = .06$ ($\text{BART}_{\text{uniform}}$), $r =$ -.03 ($\text{BART}_{\text{normal-H}}$), $r = .14$ ($\text{BART}_{\text{normal-M}}$), and $r = .05$ ($\text{BART}_{\text{normal-L}}$). Moreover, only in the $\text{BART}_{\text{normal-M}}$ 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

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

Comparison	Δ [95% HDI]	pROPE	*d*
Mean adjusted BART scores			
$BART_{normal\text{-}H}$ - $BART_{uniform}$	**4.48** [3.25, 5.70]	< .000	0.71
$BART_{normal\text{-}M}$ - $BART_{uniform}$	**2.94** [1.84, 4.06]	< .000	0.51
$BART_{normal\text{-}L}$ - $BART_{uniform}$	**7.74** [6.73, 8.73]	< .000	1.21
$BART_{normal\text{-}M}$ - $BART_{normal\text{-}H}$	**-1.54** [-2.64, -0.44]	.002	-0.27
$BART_{normal\text{-}L}$ - $BART_{normal\text{-}M}$	**4.80** [3.98, 5.63]	< .000	0.77
SD of the adjusted BART scores			
$BART_{normal\text{-}H}$ - $BART_{uniform}$	0.13, [-0.90, 1.17]	.078	-
$BART_{normal\text{-}M}$ - $BART_{uniform}$	**-1.31** [-2.37, -0.26]	.004	-
$BART_{normal\text{-}L}$ - $BART_{uniform}$	**-3.37** [-4.28, -2.42]	< .000	-
$BART_{normal\text{-}M}$ - $BART_{normal\text{-}H}$	**-1.44** [-2.50, -0.46]	.001	-
$BART_{normal\text{-}L}$ - $BART_{normal\text{-}M}$	**-2.06** [-2.95, -1.13]	< .000	-
Deviance of the adjusted BART scores from optimal behavior			
$BART_{normal\text{-}H}$ - $BART_{uniform}$	**8.47** [7.22, 9.70]	< .000	1.36
$BART_{normal\text{-}M}$ - $BART_{uniform}$	**9.94** [8.82, 11.04]	< .000	1.77
$BART_{normal\text{-}L}$ - $BART_{uniform}$	**7.73** [6.72, 8.74]	< .000	2.53
$BART_{normal\text{-}M}$ - $BART_{normal\text{-}H}$	**1.46** [0.37, 2.57]	.003	0.27
$BART_{normal\text{-}L}$ - $BART_{normal\text{-}M}$	**4.80** [3.96, 5.62]	< .000	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.

($r = .27$), 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 $BART_{normal\text{-}M}$ from the three implementations of the $BART_{normal}$ as the focus of our comparison with the $BART_{uniform}$ and report the analyses for the $BART_{normal\text{-}H}$ and $BART_{normal\text{-}L}$ in the online Supplemental Material (Section 7.1).

Compared against each other, there were some indications that the $BART_{normal\text{-}M}$ 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 $BART_{normal\text{-}M}$, 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) 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 $BART_{normal\text{-}M}$.

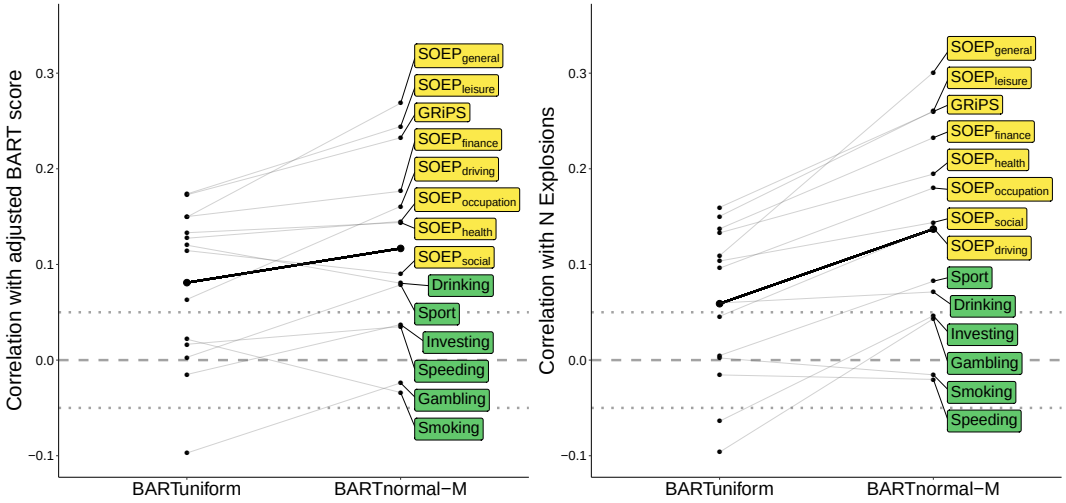

Figure 5. Convergent validity of the BART with other measures of risk taking, separately for the $\text{BART}_{\text{uniform}}$ and the $\text{BART}_{\text{normal-M}}$. The left panel shows the correlations based on the adjusted BART scores. The right panel shows the correlations based on the total number 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.

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 $\text{BART}_{\text{normal}}$ 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 $\text{BART}_{\text{normal-M}}$ 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). We thus 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 objectively 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 $\text{BART}_{\text{normal}}$, 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 $\text{BART}_{\text{normal}}$ as compared to in the $\text{BART}_{\text{uniform}}$, 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 $\text{BART}_{\text{normal}}$ as opposed to in the $\text{BART}_{\text{uniform}}$. Moreover, we expected the test–retest reliability within the $\text{BART}_{\text{normal}}$ 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 $\text{BART}_{\text{normal}}$ as compared to in the $\text{BART}_{\text{uniform}}$. Moreover, we expected the test–retest reliability within the $\text{BART}_{\text{normal}}$ 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, Lynam, Miller, & Reynolds, 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 $\text{BART}_{\text{uniform}}$, 158 completed the $\text{BART}_{\text{normal-H}}$, 157 completed the $\text{BART}_{\text{normal-M}}$, and 160 completed the $\text{BART}_{\text{normal-L}}$).

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 $\text{BART}_{\text{normal}}$ implementations were higher than those from the $\text{BART}_{\text{uniform}}$. Moreover, we compared the test–retest reliabilities within the $\text{BART}_{\text{normal}}$ 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 $\text{BART}_{\text{normal}}$ implementations were higher than that of the $\text{BART}_{\text{uniform}}$. Moreover, we compared the test–retest reliabilities within the three $\text{BART}_{\text{normal}}$ 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.% ($\text{BART}_{\text{uniform}}$), 76.3% ($\text{BART}_{\text{normal-H}}$), 76.4% ($\text{BART}_{\text{normal-M}}$), and 84.2% ($\text{BART}_{\text{normal-L}}$) of participants indicated that they believed that the explosion points were clustered, with average confidence ratings of 17.64 in the $\text{BART}_{\text{uniform}}$, 19.22 in the $\text{BART}_{\text{normal-H}}$, 19.98 in the $\text{BART}_{\text{normal-M}}$, and 25.51 in the $\text{BART}_{\text{normal-L}}$ (on a scale ranging from -50 to 50).

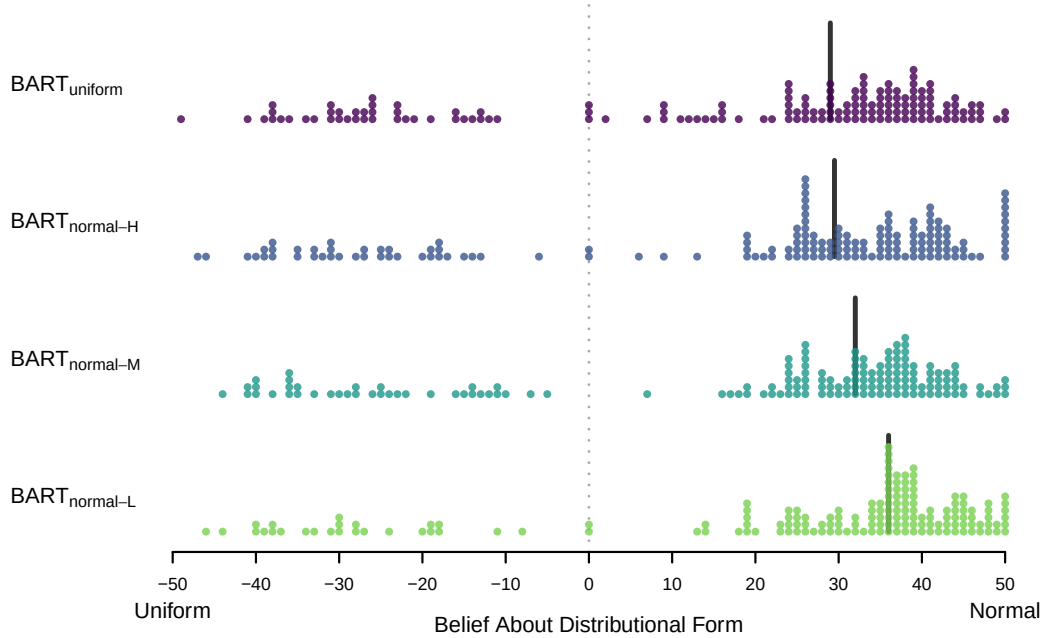

Figure 6. Distributions of participants' beliefs that the balloons' explosion points were uniformly distributed (rating of -50) vs. that they were normally distributed (rating of 50). Beliefs were assessed at the end of the task. Vertical lines indicate the median ratings, separately for the four experimental conditions. The dotted gray line indicates the center of the scale, which corresponds to minimal confidence (i.e., indifference between the two distributional forms).

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 $\text{BART}_{\text{normal}}$ became. Yet, there was no conclusive evidence for differences across the tested contrasts between the $\text{BART}_{\text{uniform}}$ and the $\text{BART}_{\text{normal}}$ implementations (b = 3.91, 95% HDI: [-1.03, 8.70], pROPE = .307, d = 0.15), the $\text{BART}_{\text{normal-M}}$ and the $\text{BART}_{\text{normal-H}}$ (b = -0.79, 95% HDI: [-6.80, 5.35], pROPE = .611, d = 0.03), and the $\text{BART}_{\text{normal-L}}$ and the $\text{BART}_{\text{normal-M}}$ (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 $\text{BART}_{\text{normal}}$ would have a higher convergent validity with other measures of risk taking than the $\text{BART}_{\text{uniform}}$. 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 ($\text{BART}_{\text{uniform}}$), r = .01 ($\text{BART}_{\text{normal-H}}$), r = .08 ($\text{BART}_{\text{normal-M}}$), and r = .05 ($\text{BART}_{\text{normal-L}}$). The total number of explosions per participant exhibited about the

same convergent validity as the adjusted BART scores, with average correlations of $r = .08$ ($BART_{uniform}$), $r = .03$ ($BART_{normal\text{-}H}$), $r = .07$ ($BART_{normal\text{-}M}$), and $r = .06$ ($BART_{normal\text{-}L}$). Moreover, only in the $BART_{normal\text{-}M}$ 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 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 $BART_{normal\text{-}M}$, 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 $BART_{normal\text{-}M}$ from the three implementations of the $BART_{normal}$ as the focus of our comparison with the $BART_{uniform}$ and report the analyses on $BART_{normal\text{-}H}$ and $BART_{normal\text{-}L}$ in the online Supplemental Material (Tables S8 and S9).

Compared to each other, there were no indications that the $BART_{normal\text{-}M}$ 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 $BART_{normal\text{-}M}$ 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 correlations 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 $r = .07$ ($BART_{uniform}$), $r = .04$ ($BART_{normal\text{-}H}$), $r = .08$ ($BART_{normal\text{-}M}$), and $r = .06$ ($BART_{normal\text{-}L}$), and those with the total number of explosions per participants were $r = .11$ ($BART_{uniform}$), $r = .05$ ($BART_{normal\text{-}H}$), $r = .10$ ($BART_{normal\text{-}M}$), and $r = .07$ ($BART_{normal\text{-}L}$).

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 $r = .40$ ($BART_{uniform}$), $r = .41$ ($BART_{normal\text{-}H}$), $r = .26$ ($BART_{normal\text{-}M}$), and $r = .43$ ($BART_{normal\text{-}L}$; 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).

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,

Table 3
Differences in Test–Retest Reliabilities of BART Indicators Between Experimental Conditions

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

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 $BART_{normal\text{-}H}$ ($r = .73$), and the $BART_{normal\text{-}M}$ ($r = .65$), and, surprisingly, somewhat lower in the $BART_{normal\text{-}L}$ ($r = .42$). There was conclusive evidence for differences in the test–retest reliabilities between the $BART_{normal\text{-}H}$ and the $BART_{normal\text{-}L}$ ($\Delta_r = .30$, 95% HDI: [.16, .45]; pROPE $< .000$), and the $BART_{normal\text{-}M}$ and the $BART_{normal\text{-}L}$ ($\Delta_r = .23$, 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 $BART_{normal\text{-}H}$ ($r = .66$) and the $BART_{normal\text{-}M}$ ($r = .63$), and somewhat lower ones in the $BART_{uniform}$ ($r = .47$), and the $BART_{normal\text{-}L}$ ($r = .48$). There was no conclusive evidence for differences 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–reliability of the $\text{BART}_{\text{normal}}$ with that of the $\text{BART}_{\text{uniform}}$. As predicted, we observed a strong mismatch between people's general task representation and the stochastic structure of the $\text{BART}_{\text{uniform}}$, and this mismatch did not emerge in the $\text{BART}_{\text{normal}}$. 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., 2012, 2011; 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., $\text{BART}_{\text{uniform}}$ vs. $\text{BART}_{\text{normal}}$), 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 $\text{BART}_{\text{normal}}$ (as compared to participants who completed the $\text{BART}_{\text{uniform}}$) 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 $\text{BART}_{\text{normal}}$ as compared to in the $\text{BART}_{\text{uniform}}$. In short, in the $\text{BART}_{\text{normal}}$ 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 systematic 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 $\text{BART}_{\text{normal}}$ 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, Simmons, Aklin, Daughters, & Dvir, 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, Dutilh, & Wagenmakers, 2011). That said, recent developments appear to mitigate the latter issue (Park, Yang, Vassileva, & Ahn, 2019), and in future work such models (and promising novel variants thereof; Pleskac & Wershbale, 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, Dhami 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, Quercia, & Willrodt, 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, Slovic, Lichtenstein, Read, & Combs, 1978; Frey, 2020; Jellison & Riskind,

1970; Loewenstein, Weber, Hsee, & Welch, 2001; Morrongiello & Lasenby-Lessard, 2007; Phillips, Hertwig, Kareev, & Avrahami, 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 between 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 (Dhami 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 $\text{BART}_{\text{normal}}$. 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 $\text{BART}_{\text{normal}}$ 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, Hampton, Biers, & Kirton, 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

[7] Task 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, Hancock, Arthur, & Chrysler, 1993; Roccio, 1995).

[8] Note 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.

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, Pachur, Hertwig, & Pleskac, 2018; Pleskac et al., 2020; Pleskac & Hertwig, 2014; Schonberg et al., 2011). The absence of substantial improvements in the BART's external validity (i.e., in response to the elementary stochastic adaptions 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 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, Sonnentag, Niessen, & Zapf, 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, Arendasy, Sommer, & Bognar, 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.

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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1037/1064-1297.13.4.311

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (Second ed.). Berkley, CA: 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. doi: 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. doi: 10.1016/j.jebo.2012.12.023

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*(6), 959–988. doi: 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. doi: 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). Retrieved from `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. doi: 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. doi: 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. doi: 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. Retrieved from `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. doi: 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. doi: 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. doi: 10.1037/a0038118

Gibson, J. J. (1986). *The ecological approach to visual perception.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). *Rstanarm: Bayesian applied regression modeling via Stan (Version 2.17.4).* Retrieved from `http://mc-stan.org/`

Gray, J. A. (1982). On mapping anxiety. *Behavioral and Brain Sciences*, *5*(3), 506–534. doi: 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*. doi: 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). New York: 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. doi: 10.1177/0269881114550352

Helfinstein, S. M., Schonberg, T., Congdon, E., Karlsgodt, K. H., Mumford, J. A., Sabb, F. W., ... Poldrack, R. A. (2014). Predicting risky choices from brain activity patterns. *Proceedings of the National Academy of Sciences*, *111*(7), 2470–2475. doi: 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. doi:

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). New York: Cambridge University Press.

Holt, C. A., & Laury, S. (2002). Risk aversion and incentive effects. *The American Economic Review*, *92*(5), 1644–1655. doi: 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. doi: 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. doi: 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. doi: 10.1037/h0029601

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. doi: 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. doi: 10.1207/s15324818ame1703_1

Knight, F. H. (1921). *Risk, uncertainty, and profit.* New York: 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. doi: 10.1080/00140139308967941

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi: 10.1037/a0029146

Kruschke, J. K., & Meredith, M. (2018). *BEST: Bayesian estimation supersedes the t-test (Version 0.5.1).* Retrieved from `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. doi: 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. doi: 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. doi: 10.1037/1064-1297.11.1.26

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. doi: 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.,

... 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. doi: 10.1037/1076-898X.8.2.75

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. doi: 10.1016/j.addbeh.2004.02.035

Leuker, C., Pachur, T., Hertwig, R., & Pleskac, T. J. (2018). Exploiting risk–reward structures in decision making under uncertainty. *Cognition*, *175*, 186–200. doi: 10.1016/j.cognition.2018.02.019

Li, X., Pan, Y., Fang, Z., Lei, H., Zhang, X., Shi, H., ... Rao, H. (2019). Test-retest reliability of brain responses to risk-taking during the balloon analogue risk task. *NeuroImage*, 116495. doi: 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. doi: 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. doi: 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. doi: 10.2307/2232669

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, 1–8. doi: 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. doi: 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 tisk task (BART). *PloS one*, *7*(6), e39135. doi: 10.1371/journal.pone.0039135

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. doi: 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*, 1–15. doi: 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. doi: 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).* Retrieved from `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. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 38, pp. 95–99). Santa Monica, CA. doi: 10.1177/154193129403800118

Morrongiello, 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. doi: 10.1136/ip.2005.011296

Mousavi, S., & Gigerenzer, G. (2014). Risk, uncertainty, and heuristics. *Journal of Business Research*, *67*(8), 1671–1678. doi: 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. doi: 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*. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1037/a0014245

Pleskac, T. J., & Wershbale, A. (2014). Making assessments while taking repeated risks: A pattern of multiple response pathways. *Journal of Experimental Psychology: General*, *143*(1), 142–162. doi: 10.1037/a0031106

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from `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. doi: 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* (Vol. 2, pp. 122–181). Mahwah, NJ: 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. doi: 10.1027/1015-5759/a000335

Schonberg, T., Fox, C. R., Mumford, J. A., Congdon, E., Trepel, 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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*. doi: 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). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Tisdall, L., Frey, R., Horn, A., Ostwald, D., Horvath, L., Blankenburg, F., ... Mata, R. (2020). Brain-behavior associations for risk taking depend on the measures used to capture individual differences. *Frontiers in Behavioral Neuroscience*, *14*, 587152. doi: 10.3389/fnbeh.2020.587152

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151–176. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1002/bdm.2102

Appendix C: Steiner & Grieder (2020)

Steiner, M. D. & Grieder, S. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53), 2521. doi: 10.21105/joss.02521

The Journal of Open Source Software

EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools

Markus D. Steiner[1] **and Silvia Grieder**[2]

1 Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel, Switzerland **2** Division of Developmental and Personality Psychology, Department of Psychology, University of Basel, Switzerland

DOI: 10.21105/joss.02521

Software

- Review
- Repository
- Archive

Editor: Frederick Boehm

Reviewers:

- @jacobsoj
- @chainsawriot

Submitted: 19 July 2020
Published: 16 September 2020

Summary

In the social sciences, factor analysis is a widely used tool to identify latent constructs underlying task performance or the answers to questionnaire items. Exploratory factor analysis (EFA) is a data-driven approach to factor analysis and is used to extract a smaller number of common factors that represent or explain the common variance of a larger set of manifest variables (see, e.g., Watkins, 2018 for an overview). Several decisions have to be made in advance when performing an EFA, including the number of factors to extract, and the extraction and rotation method to be used. After a factor solution has been found, it is useful to subject the resulting factor solution to an orthogonalization procedure to achieve a hierarchical factor solution with one general and several specific factors. This situation especially applies to data structures in the field of intelligence research where usually high, positive factor intercorrelations occur. From this orthogonalized, hierarchical solution, the variance can then be partitioned to estimate the relative importance of the general versus the specific factors using omega reliability coefficients (e.g., McDonald, 1999).

EFAtools is an R package (R Core Team, 2020) that enables fast and flexible analyses in an EFA framework, from tests for suitability of the data for factor analysis and factor retention criteria to hierarchical factor analysis with Schmid-Leiman transformation (Schmid & Leiman, 1957) and McDonald's omegas (e.g., McDonald, 1999). The package's core functionalities are listed in Table 1.

Statement of Need

Compared to other R packages with which EFA can be performed, *EFAtools* has several advantages, including fast implementations using *Rcpp* (Eddelbuettel & Balamuta, 2017; Eddelbuettel & Sanderson, 2014), more flexibility in the adjustment of implementation features, the ability to reproduce the R *psych* (Revelle, 2020) and SPSS (IBM, 2015) implementations of some analyses methods (see vignette *Replicate SPSS and R psych results with EFAtools*), as well as the inclusion of recommended implementations for these methods based on simulation analyses (Grieder & Steiner, 2020). Finally, the package includes the implementation of the, as of yet, most comprehensive set of factor retention criteria in R, including recently developed criteria such as the Hull method (Lorenzo-Seva, Timmerman, & Kiers, 2011), comparison data (Ruscio & Roche, 2012), and the empirical Kaiser criterion (Braeken & van Assen, 2016). As recommended by Auerswald & Moshagen (2019), multiple factor retention criteria should be examined simultaneously to check their convergence, which now is easily possible with a comprehensive function in *EFAtools* incorporating all implemented factor retention criteria for simultaneous application. Minor advantages over and above the existing implementations in R

include that when intending to perform a Schmid-Leiman transformation, this can be done on an obliquely rotated solution obtained with functions from the *EFAtools* or the *psych* package instead of being forced to perform the whole EFA procedure again. Moreover, our implementation of McDonald's omegas calculations include the possibility of manual variable-to-factor correspondences (as are needed for variance partitioning for predetermined / theoretical composites) in addition to automatically determined variable-to-factor correspondences (as done, for example, in the *psych* package). Further, the *EFAtools* function to compute McDonald's omegas can easily be applied on *EFAtools* and *psych* Schmid-Leiman solutions as well as on *lavaan* (Rosseel, 2012) second-order, bifactor, and single factor solutions (including solutions from multiple group analyses).

Development and Purpose

EFAtools was designed for use in the social sciences in general and is especially suitable for research on cognitive abilities or other hierarchically organized constructs as well as for more time-consuming applications such as in simulation analyses. Its development arose from the need for a tool for easy replication and comparison of EFA solutions from different programs, namely R and SPSS (Grieder & Steiner, 2020), and has already been used in another publication (Grieder & Grob, 2019). The package was then expanded for a broader, easy, fast, and flexible use of EFA tools such that it is now suitable for most projects within the EFA framework.

Table 1: Core functionalities of *EFAtools*.

Topic	Method	Function
Suitability for factor analysis	Bartlett's test of sphericity	`BARTLETT()`
	Kaiser-Meyer-Olkin criterion	`KMO()`
Factor retention criteria	Comparison data	`CD()`
	Empirical Kaiser criterion	`EKC()`
	Hull method	`HULL()`
	Kaiser-Guttman criterion	`KGC()`
	Parallel analysis	`PARALLEL()`
	Scree plot	`SCREE()`
	Sequential model tests	`SMT()`
	RMSEA lower bound criterion	`SMT()`
	AIC criterion	`SMT()`
Factor extraction methods	Principal axis factoring	`EFA()`
	Maximum likelihood	`EFA()`
	Unweighted least squares	`EFA()`
Rotation methods	Orthogonal: Varimax, equamax, quartimax, geominT, bentlerT, bifactorT	`EFA()`
	Oblique: Promax, oblimin, quartimin, simplimax, bentlerQ, geominQ, bifactorQ	`EFA()`
Factor scores	Different methods for calculating factor scores	`FACTOR_SCORES()`
Hierarchical factor analysis	Schmid-Leiman transformation	`SL()`
	McDonald's omegas	`OMEGA()`

Note. All functions for suitability for factor analysis and factor retention criteria can be called in any desired combination using the `N_FACTORS()` function.

The Journal of Open Source Software

Installation

The *EFAtools* package can be installed from CRAN using `install.packages("EFAtools")`. Moreover, the development version can be installed from GitHub (https://github.com/mdsteiner/EFAtools) using `devtools::install_github("mdsteiner/EFAtools", build_vignettes = TRUE)`.

Acknowledgements

We thank Dirk Wulff for helpful suggestions concerning the C++ implementations.

References

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. doi:10.1037/met0000200

Braeken, J., & van Assen, M. A. L. M. (2016). An empirical Kaiser criterion. *Psychological Methods*, *22*(3), 450–466. doi:10.1037/met0000074

Eddelbuettel, D., & Balamuta, J. J. (2017). Extending R with C++: A brief introduction to Rcpp. *PeerJ Preprints*, *5*, e3188v1. doi:10.7287/peerj.preprints.3188v1

Eddelbuettel, D., & Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, *71*, 1054–1063. doi:10.1016/j.csda.2013.02.005

Grieder, S., & Grob, A. (2019). Exploratory factor analysis of the Intelligence and Development Scales–2: Implications for theory and practice. *Assessment*. Advance online publication. doi:10.1177/1073191119845051

Grieder, S., & Steiner, M. D. (2020). *Algorithmic jingle jungle: A comparison of implementations of Principal Axis Factoring and promax rotation in R and SPSS*. PsyArXiv. doi:10.31234/osf.io/7hwrm

IBM, C. (2015). *IBM SPSS Statistics for Macintosh, Version 23.0*. Armonk, NY.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The Hull Method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364. doi:10.1080/00273171.2011.564527

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum. ISBN: 978-1-4106-0108-7

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:10.18637/jss.v048.i02

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*(2), 282–292. doi:10.1037/a0025697

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61. doi:10.1007/BF02289209

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*, 219–246. doi:10.1177/0095798418771807

Appendix D: Grieder & Steiner (2020)

Grieder, S. & Steiner, M. D. (2020)[11]. *Algorithmic jingle jungle: A comparison of implementations of principal axis factoring and promax rotation in R and SPSS.* Manuscript submitted for publication. Preprint doi: 10.31234/osf.io/7hwrm

[11]Shared first authorship.

Algorithmic Jingle Jungle: A Comparison of Implementations of Principal Axis Factoring and Promax Rotation in R and SPSS

Silvia Grieder*, Markus D. Steiner*
University of Basel

Abstract

A statistical procedure is assumed to produce comparable results across programs. Using the case of an exploratory factor analysis procedure—principal axis factoring (PAF) and promax rotation—we show that this assumption is not always justified. Procedures with equal names are sometimes implemented differently across programs: a jingle fallacy. Focusing on two popular statistical analysis programs, we indeed discovered a jingle jungle for the above procedure: Both PAF and promax rotation are implemented differently in the *psych* R package and in SPSS. Based on analyses with 247 real and 216,000 simulated data sets implementing 108 different data structures, we show that these differences in implementations can result in fairly different factor solutions for a variety of different data structures. Differences in the solutions for real data sets ranged from negligible to very large, with 38% displaying at least one different indicator-to-factor correspondence. A simulation study revealed systematic differences in accuracies between different implementations, and large variation between data structures, with small numbers of indicators per factor, high factor intercorrelations, and weak factors resulting in the lowest accuracies. Moreover, although there was no single combination of settings that was superior for all data structures, we identified implementations of PAF and promax that maximize performance on average. We recommend researchers to use these implementations as best way through the jungle, discuss model averaging as a potential alternative, and highlight the importance of adhering to best practices of scale construction.

Keywords: software comparison, exploratory factor analysis, principal axis factoring, promax rotation

Psychological research is mainly conducted using quantitative methods. Whereas in the early days of psychology statistical procedures had to be implemented by hand, today a variety of programs exists for this purpose. Generally, implementations of statistical procedures are thought to produce equivalent results

across programs; at least their interchangeable use in scientific publications suggests as much. However, to date, detailed comparisons of these implementations (i.e., on a code or output level) are scarce (for exceptions, see, e.g., Hodges, Stone, Johnson, Carter, & Lindsey, 2020; Kotenko, 2017; Stanley, 2015), and it is thus unclear whether this interchangeable use of programs is justified (for meta-level comparisons, see, e.g., Gosh, 2019; Klinke & Härdle, 2010; MacCallum, 1983; Marr-Lyon, Gupchup, & Anderson, 2012).

Our attention was drawn to this topic in the course of the review process of Grieder and Grob (2020), where the authors conducted exploratory factor analyses (EFA). A reviewer of Grieder and Grob (2020) suggested testing the robustness of results using a second program. It became apparent that although the same statistical methods were applied, results were not comparable between these programs, and even led to different interpretations and conclusions. As it turned out, this case was no exception (e.g., Collins, 2016; del Rio, 2017; GaryStats, 2017; Hodges et al., 2020; krissen, 2018; u/kriesniem, 2018, see also Newsom, 2020 on EFA). As Ershova and Schneider (2018) point out, results from a specific statistical method can even differ between different versions of the same program if the implementation of algorithms are changed, which often happens without explicit notification of the users[1]. Given the instances referenced above, it seems likely that most people assume implementations of the same method in different programs to yield equivalent—or at least highly comparable—results. As a consequence, a researcher reporting results based on one program might be criticized if these results are not reproducible by another researcher employing another program. This might become more of an issue with increasing popularity and promotion of the open science movement (see Gernsbacher, 2018), as data may be shared more often, and it might also contribute to one of the most pressing contemporary issues in psychology—the replication crisis (Ershova & Schneider, 2018; Munafò et al., 2017; Open Science Collaboration, 2015), for example when it comes to ongoing construct validations (see Flake, Pek, & Hehman, 2017) or direct replication studies (Hodges et al., 2020). In factor analysis, the worst consequence of differences in implementations could be a misalignment of which indicator was classified to be part of which latent construct across results obtained from different implementations. This is exactly what happened in the personal example mentioned above, and it might seemingly yield evidence against the validity of a scale

[1]Such changes are often listed in a changelog or news file, but this still requires users to actively consult the changelog after every software update.

---

*Both authors contributed equally to this work.

Silvia Grieder, Division of Developmental and Personality Psychology, Department of Psychology, University of Basel; Markus D. Steiner, Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel.

We thank Rui Mata, Dirk Wulff, Renato Frey, Alexandra Bagaïni, and the whole CDS team, as well as Gary L. Canivez for valuable comments on an earlier version of this manuscript. We thank Laura Wiles for proofreading.

Corresponding author: Silvia Grieder, Department of Psychology, University of Basel, Missionsstrasse 62, 4055 Basel, Switzerland. E-mail: silvia.grieder@unibas.ch

in ongoing scale validation (Flake et al., 2017), even though the differences might be due to the implementation of the same statistical procedure used.

The case where different concepts (in our case: different implementations of a statistical procedure) are referred to by the same name is known as the jingle fallacy (Thorndike, 1904). The present work provides a first step to gauging the extent to which a "jingle jungle" exists in the implementations of some statistical procedures across different programs. Moreover, we map this jungle by revealing different ways through it and gauge the resulting implications by looking at the size of possible differences in results. Finally, we try to navigate through the jungle by identifying the implementation that renders the most accurate results. In the present study, we focus on implementations of a specific procedure within a frequently used statistical framework—EFA—in two of the most often used programs for statistical analyses in psychological research (Dunn, 2011): *SPSS* (IBM Corp., 2020) and *R* (R Core Team, 2020).

Exploratory Factor Analysis

Factor analysis is a widely used tool to identify latent constructs underlying task performance or responses to questionnaire items. In factor analysis, the variance in a larger number of variables or indicators is sought to be accounted for by a smaller number of latent factors. A data-driven approach to factor analysis is EFA, which was originally developed by Spearman (1904, 1927) as a method to extract a common factor—a mathematical entity that accounts for the interrelations of test scores from different cognitive tasks (i.e., for the positive manifold of cognitive performances). This common entity, the general factor, is the construct thought to underlie manifest variables, such as subtest scores from intelligence tests. In EFA, intercorrelations between a given set of indicators are analyzed and a smaller number of factors is extracted that explain a maximum of the common variance between these indicators.

Two crucial decisions have to be made in advance when performing an EFA. First, the number of factors to retain needs to be determined. It is recommended to use multiple retention criteria for this purpose. Auerswald and Moshagen (2019), for example, suggest to use sequential χ^2 model tests in combination with either parallel analysis (PA; Horn, 1965), the empirical Kaiser criterion (Braeken & Van Assen, 2017), or the Hull method (Lorenzo-Seva, Timmerman, & Kiers, 2011). In addition to quantitative criteria, qualitative criteria, such as theoretical considerations and the plausibility of the factor solution (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Hayton, Allen, & Scarpello, 2004; Watkins, 2018), should also be considered.

Second, the factor extraction method and the rotation method used to seek *simple structure*—that is, a solution where each indicator loads substantially onto one, and only one, factor—need to be chosen. One of the most commonly used factor extraction methods is iterative principal axis factoring (PAF). Compared to another frequently used and recommended method, maximum likelihood estimation (ML), PAF has several advantages. First, it has no distributional assumptions, whereas ML requires the data to follow a multivariate normal distribution (e.g., Fabrigar et al., 1999). Second, it is more robust in the case of unequal factor loadings, few indicators

per factor, and small sample sizes (Briggs & MacCallum, 2003; De Winter & Dodou, 2012). Finally, it is better able to recover weak factors (Briggs & MacCallum, 2003; De Winter & Dodou, 2012). This means that PAF is less likely to produce Heywood cases (that is, negative variances/standardized variances ≥ 1) and non-convergence in the aforementioned data structures compared to ML (Briggs & MacCallum, 2003; De Winter & Dodou, 2012; Fabrigar et al., 1999). On the other hand, ML enables the computation of multiple fit indices and of significance tests for the factor loadings (e.g., Fabrigar et al., 1999). In our study, we focus on PAF as a factor extraction method.

Once the specified number of common factors is extracted, the solution is typically rotated in an attempt to obtain simple structure. Rotations can be categorized into orthogonal and oblique rotations. In an orthogonally rotated solution, the resulting factors are uncorrelated. The most popular orthogonal rotation method is varimax (Kaiser, 1958; Watkins, 2018). In most cases, however, the resulting factors are assumed to be at least somewhat correlated and thus an oblique rotation is more appropriate. Many oblique rotation procedures also start with an orthogonal rotation, but then the orthogonality constraints are lessened and the factors are allowed to correlate. In the case of correlated factors, this results in a solution that approaches simple structure even further than an orthogonal solution. If the factors are uncorrelated in reality, an oblique rotation will produce orthogonal factors as well. This is why oblique rotations are generally recommended over orthogonal rotations (Fabrigar et al., 1999; Gorsuch, 1983; Watkins, 2018). The most popular oblique rotation methods are promax (Hendrickson & White, 1964) and oblimin (Carroll, 1958; Jennrich & Sampson, 1966; Watkins, 2018). If an oblique factor solution contains substantial factor intercorrelations, this implies that a hierarchical structure is present (i.e., one or more constructs can be assumed that explain these intercorrelations). In this case, it is useful to make the hierarchical nature explicit by transforming the oblique solution accordingly, for example using the Schmid-Leiman transformation (Schmid & Leiman, 1957). In this article, we focus on oblique factor solutions after promax rotation, as this can be used with both hierarchical and nonhierarchical data structures. The implementations of PAF and promax rotation in R and SPSS are described below.

Present Study

The major aim of this study is to compare implementations and results of a commonly used EFA procedure—PAF and promax rotation—in R, version 4.0.3 (R Core Team, 2020) and in SPSS, version 27 (IBM Corp., 2020)[2]. In R, we used the *psych* package, version 2.0.12. (Revelle, 2020; henceforth referred to as R psych), a very popular and extensive R package that has also influenced EFA implementations in python (Biggs, 2019) and has been recommended to be used with the R plugin in

[2]Our results are also valid for SPSS version 23, and probably versions 24 to 26, as the algorithms for the investigated procedures did not change from version 23 through 27 according to the algorithms manuals (IBM Corp., 2014, 2016, 2017) and we found results to be identical between versions 23 and 27 for these procedures.

SPSS for some calculations (IBM Support, 2020). We first compared the implementations in R psych and SPSS on a code-level. As the source code for SPSS is not publicly available, we relied on the algorithms manual in which the formulas of the implementations are provided (IBM Corp., 2017). As a next step, we were interested to see how the identified differences between the two implementations would impact results. To this end, we compared results from the two implementations for a large set of real data. Finally, we wanted to see whether there is an implementation of PAF and promax rotation that renders more accurate results in general, or at least for certain data structures. To this end, we compared the ability to recover a set of true population models across all possible combinations of the considered settings for PAF and promax rotation—including the two combinations used in R psych and SPSS—in a simulation study based on a large set of population models (varying, e.g., factor intercorrelations and the number of indicators per factor). This also allowed us to determine whether the implementation or the data structure is more important for accurate results. Thus, our goal was to explore whether a jingle jungle indeed exists for the investigated procedure, try to map it, and seek the best way through it.

To facilitate comparisons, we first ran analyses in the programs mentioned above and then reproduced results from both programs using our own functions included in a dedicated R package—*EFAtools* (Steiner & Grieder, 2020). We then conducted all further analyses with our own functions that enable a flexible use of all combinations of settings needed for the simulation analyses and that are faster due to C++ implementations of the iterative procedures. Results on how well our functions reproduced the original implementations in R psych and SPSS are provided below. Supplemental material (SM) to this study, as well as all analysis scripts and many of the data sets used for the real data analyses, are available at `https://osf.io/a836q`.

Implementations in R psych and SPSS

Principal Axis Factoring

PAF is a least squares fitting approach in EFA. It uses the variances and covariances of a given set of indicators to reduce dimensionality by extracting a prespecified number of factors such that they explain a maximum of the common variance in these indicators (often, a correlation matrix is used to this end). The standard way of performing PAF in R is with the `fa` function in the psych package, and in SPSS with the `FACTOR` algorithm. We now briefly describe the differences in PAF implementations in the two programs; these are also listed in Table A1. A more detailed description of PAF and its implementation in R psych and SPSS is included in the SM (section 1).

In a first step, both in R psych and SPSS, the correlation matrix is tested for different properties. If the correlation matrix is not positive definite, R psych will perform smoothing to produce a highly similar positive definite matrix to proceed with. In contrast, SPSS will throw an error and abort if a non-positive definite matrix is entered.

In a next step, initial communalities[3] are estimated and used to replace the diagonal of the correlation matrix. Several approaches exist for deriving initial estimates; three of them being (a) *unity* (i.e., each initial communality is set to one, thus the correlation matrix remains unchanged), (b) the *maximum absolute correlation* (MAC) of an indicator with any other indicator, and (c) the *squared multiple correlations* (SMCs; Gorsuch, 1983; Harman, 1976). It has been advocated to rely on SMCs as initial estimates (Dwyer, 1939; Guttman, 1956; Roff, 1936; Wrigley, 1957)—which is what both the R psych and SPSS implementations do by default. When SMCs are entered into the diagonal of the correlation matrix, it is often the case that the matrix is no longer positive semidefinite; that is, some of its eigenvalues are negative. The PAF procedure, however, involves taking the square root of the m largest eigenvalues and therefore cannot be executed if any of these m eigenvalues—where m is the number of factors to extract—are negative. When SMCs cannot be used, R psych suggests using unity as initial communality estimates, which may lead to convergence, but may also lead to inflated final communality estimates (Gorsuch, 1983). It has therefore been recommended to use MACs instead of unity when SMCs fail (Gorsuch, 1983)—which is what SPSS supposedly does (IBM Corp., 2017). However, using SMCs rarely fails in SPSS, as SPSS takes the absolute of the eigenvalues, thereby avoiding negative eigenvalues during the iterative PAF procedure (IBM Corp., 2017). In R psych, using SMCs will fail whenever any of the m largest eigenvalues are negative. Thus, R psych and SPSS deal differently with negative eigenvalues, which results in different cases where using SMCs will fail.

After the initial communality estimates have been determined, the final communalities are estimated in an iterative process (see SM section 1 for details). This process is continued until an arbitrary convergence criterion is reached, which, by default, is 10^{-3} for both R psych and SPSS. Others have suggested more strict criteria, such as 10^{-5} (Mulaik, 2010) or 10^{-6} (Briggs & MacCallum, 2003). When testing convergence—that is, testing whether the differences from one iteration to the next are small enough and thus the current solution is considered stable—SPSS tests against the maximum difference in any single communality estimate, whereas R psych tests against the difference in the sum of all communalities.

To summarize, we found three differences between the R psych and SPSS PAF implementations, namely not taking versus taking the absolute value of eigenvalues, using unity versus MACs as initial communality estimates if SMCs fail, and using different referents when testing convergence.

Promax Rotation

Once the specified number of factors is extracted, a rotation is typically performed to achieve an interpretable solution. Promax is a fast and efficient method for oblique factor rotation. In this procedure, a varimax rotation (usually preceded by Kaiser normalization; Kaiser, 1958) is performed first to obtain an orthogonal

[3]Communality refers to the proportion of variance of an indicator that can be accounted for by the common factors (Gorsuch, 1983) and thus provides an indication of how strongly the indicator is related to all other considered indicators.

solution, which is then transformed into an oblique solution (Hendrickson & White, 1964). Here, we again only briefly introduce the differences in promax implementations (see Table A1 for a summary). A more detailed description of promax and its implementation in R psych and SPSS is included in the SM (section 1).

In SPSS, promax with Kaiser normalization is implemented as a rotation method in the `FACTOR` algorithm. In R, there are at least two functions available to perform a promax: the `promax` function in the *stats* package (R Core Team, 2020) and the `Promax` function in the *psych* package, both enabling promax rotation with and without Kaiser normalization. For comparability with SPSS, we used the promax implementation with Kaiser normalization called in the R psych `fa` function.

As stated above, a varimax rotation is performed first in the promax procedure. This rotation is implemented differently in R psych and SPSS. In SPSS, the original varimax procedure from Kaiser (1958) is implemented (IBM Corp., 2017). However, the varimax criterion seems to be slightly different from the original one (see SM, section 1 for the original version and the adapted version implemented in *EFAtools*). The `varimax` function called in R psych instead uses singular value decomposition for the rotation and the sum of the singular values as varimax criterion (see also Jennrich, 2001).

Between the implementations of the subsequent steps of promax, we found only one more difference. While R psych exactly follows the original promax procedure reported in Hendrickson and White (1964), SPSS deviates from the original procedure in that it performs a row normalization of the target matrix from the varimax solution in the first step of the promax procedure (see SM, section 1 for details). Cureton (1976) provides some evidence for promax with row normalization to outperform unnormalized (i.e., original) promax. However, there exists no further evidence for or against row normalization. In most studies on promax, either both versions or the original, unnormalized version from Hendrickson and White (1964) were used (e.g., Jennrich, 2006; Lorenzo-Seva, 1999; Tataryn, Wood, & Gorsuch, 1999).

In promax rotation, the elements of the target matrix from the varimax rotation are raised to a power k. Initial evidence suggested that $k = 4$ leads to the most accurate results (Cureton, 1976; Hendrickson & White, 1964), and both R psych and SPSS use this value by default. In contrast, more recent evidence based on a large scale Monte Carlo simulation study showed that $k = 3$ is preferable in most cases for unnormalized promax, and $k = 2$ is preferable for normalized promax (Tataryn et al., 1999).

To summarize, we found two differences between the R psych and SPSS promax implementations, namely the type of the varimax rotation (original versus singular value decomposition) and the use of an unnormalized versus a row-normalized target matrix.

We thus indeed discovered a jingle jungle regarding the implementations of the same statistical procedure in two different programs and have begun to map it by pointing out the differences in the implementations. Next, we were interested in estimating the impact thereof on the resulting factor solutions. To be able to do so in the most efficient way, we reproduced the implementations with our own functions included in a dedicated R package (Steiner & Grieder, 2020).

Reproduction with the EFAtools package

To facilitate comparisons between the two, we reproduced both the R psych and SPSS implementations in the *EFAtools* package in R (Steiner & Grieder, 2020). This enabled fast comparisons for multiple data sets in the same program.

We tested how well our EFA function was able to reproduce the R psych and SPSS implementations using real and simulated data sets. As real data sets, we used four correlation matrices also included in the real data analyses reported below. Specifically, these included data on the Domain-Specific Risk-Taking scale (DOSPERT), the Intelligence and Development Scales–2 (IDS-2), and the Woodcock-Johnson IV (WJIV) on 3- to 5- and 20- to 39-year-olds (see Table *Real_data_description.xlsx* in the online repository for descriptions, `https://osf.io/pcrqu/`). As simulated data sets, we used correlation matrices derived from four selected population models constructed for the simulation analyses reported below: Case 18|3|6, case 6|3|6, case 18|3|46|3c, and case 18|6|369wb, all with strong factor intercorrelations (see Table A2 for an overview of the cases, and SM, sections 5 and 6 for the detailed models). For the pattern matrices of these population models, we apply the following naming convention throughout the manuscript: the population pattern matrices are indicated with a code in the form $p|m|\boldsymbol{\lambda}$, where p is the number of indicators, m the true number of factors, and $\boldsymbol{\lambda}$ the set of unique non-zero pattern coefficients without the period (e.g., Case 18|6|369wb is a pattern matrix with 18 indicators, six factors, and non-zero pattern coefficients of .3, .6, and .9, mixed within and between factors). If cross-loadings are present, their number is indicated in a fourth compartment, as in case 18|3|46|3c, where 3 cross-loadings are present.

With the EFAtools package, we were able to reproduce all unrotated PAF loadings, varimax loadings, and pattern coefficients from a promax rotation from the R psych `fa` function to at least the 14th decimal (see Table S1 for detailed results). From the SPSS `FACTOR` algorithm, we were able to reproduce unrotated PAF loadings to at least the 9th decimal, and both varimax loadings and pattern coefficients from a promax rotation to at least the 4th decimal (see Table S1 for detailed results).

Differences between the R psych and SPSS Solutions for Real Data Sets

Our previous analyses show that a jingle jungle does indeed exist for the implementations of PAF and promax in R psych and SPSS. But what impact does this have on the resulting factor solutions? As a first means to answer this question, we used correlation matrices from a heterogeneous collection of real data sets, mainly on cognitive abilities. These data sets varied with respect to the number of indicators, number of proposed first-order factors, and sample characteristics such as sample size, age, sex, socioeconomic status, and health status.

Methods

Data Sets. We analyzed a total of 247 correlation matrices, of which 219 were on cognitive abilities, 19 on personality, six on risk taking, and three on health and physical variables. Sample sizes varied between 22 and 619,150 (Mdn = 180), the

number of indicators varied between 6 and 300 (Mdn = 16), the proposed number of first-order factors varied between 1 and 45 (Mdn = 4), the indicator-to-factor ratio varied between 2.0 and 34.0 (Mdn = 4.2), and the sample size-to-indicator ratio varied between 1.2 and 5,159.6 (Mdn = 9.3). For more information on the data sets and their sources, see Table *Real_data_description.xlsx* in the online repository (`https://osf.io/pcrqu/`).

Statistical Analyses. As a first step, we determined the number of factors to extract. To this end, for each data set, we first determined the proposed number of factors from the literature and second, performed a PA based on SMCs with 1,000 simulated data sets[4]. Based on recommendations by Crawford et al. (2010), we tested the empirical eigenvalues of the data against the 95th percentile of the random eigenvalues for the first factor and against the mean random eigenvalues for subsequent factors.

As is commonly done, we then used the larger of the two numbers of factors—theoretical or data-driven—as an initial number of factors to extract for PAF with promax rotation. If no admissible solution was found with this initial number of factors after promax rotation, the number of factors was reduced by one and the procedure was repeated, and so on, until an admissible solution was achieved with both implementations (R psych and SPSS). A solution was deemed admissible if there were no Heywood cases (defined as communalities or pattern coefficients $\geq$ .998) and each factor displayed at least two salient pattern coefficients (i.e., $\geq$ .30; Gorsuch, 1983; Kline, 1997)[5].

We recorded the number of factors for the final promax–rotated solution that worked for both implementations and the number of factors for the first admissible solution for both the R psych and SPSS implementations, to gauge how frequently these differed. Then, we determined the frequencies of differences in indicator-to-factor correspondences between the final solutions from R psych and SPSS. A different indicator-to-factor correspondence occurs if the same indicator loads saliently onto different factors in the two solutions, or if it only displays a salient loading in one solution, but not in the other. Hence, differences in indicator-to-factor correspondences are also possible for one-factor solutions.

Moreover, we examined overall, mean, and maximum differences in loadings/-pattern coefficients after PAF without rotation, PAF with varimax rotation, and PAF with promax rotation for each data set. By separating these three analysis steps, we were able to determine which led to the largest differences in results. In addition to differences in loadings, we also examined overall, mean, and maximum factor congruence for these three analysis steps. Factor congruence is an indicator of the similarity between factors that ranges from -1 to 1, with higher values indicating higher simi-

[4]This factor retention method is available in both R (packages *EFAtools*; *nFactors*, Raiche & Magis, 2020; *paran*, Dinno, 2018; *hornpa*, Huang, 2015; and *psych*) and SPSS (with syntax provided by O'Connor, 2000, updated version available at `https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html`)

[5]These admissibility criteria were administered automatically. Due to the large number of data sets, it was not possible to manually inspect the factor solutions for plausibility, as is usually done in EFA

larity (Burt, 1948). Values between .85 and .94 are interpreted as fair similarity, and values higher than .95 as good similarity (Lorenzo-Seva & Ten Berge, 2006).

Finally, we investigated a possible relationship of the mean and maximum average differences in pattern coefficients from the final promax-rotated solutions with the indicator-to-factor ratio, as this ratio has been shown to affect the accuracy of a factor solution (MacCallum, Widaman, Zhang, & Hong, 1999). To achieve this, we performed Bayesian gamma regression analyses with a log-link function using the *rstanarm* R package (Goodrich, Gabry, Ali, & Brilleman, 2018) with default priors and the *bayestestR* package, version 0.7.2 (Makowski, Ben-Shachar, & Lüdecke, 2019). We determined credibility with the 95% highest density interval (HDI; Kruschke, 2018).

Results

Of the 247 data sets subjected to PAF with promax rotation, 31 produced non-admissible solutions only and were therefore not included in further analyses. Of the 216 data sets for which final admissible solutions were found in both R psych and SPSS, the number of factors for the first admissible solution in R psych and SPSS differed in 22.7% of the data sets, with the solution from one implementation displaying up to four factors more than the solution from the other implementation. Of this subset of data sets, the first admissible solution for SPSS more often showed a higher number of factors compared to R psych (61.2%) than vice versa (38.8%). The final solutions on which all further comparisons between the two implementations were based were all achieved using SMCs as initial communalities (i.e., the m largest eigenvalues were positive for all solutions).

Figure 1 shows the distributions of the overall, mean, and maximum absolute differences in loadings/pattern coefficients as well as the overall, mean, and minimum factor congruence between the R psych and SPSS solutions. The differences in unrotated PAF loadings ranged overall between 0.00 and 0.24, the average mean difference per solution (M_{mean}) was 0.00, and the average maximum difference per solution (M_{max}) was 0.01. After varimax rotation, the differences in loadings ranged overall between 0.00 and 0.28 ($M_{mean} = 0.00$, $M_{max} = 0.01$), and after promax rotation, the differences in pattern coefficients ranged overall between 0.00 and 0.65 ($M_{mean} = 0.02$, $M_{max} = 0.11$). The factor congruence for the unrotated PAF solutions ranged overall from .20 to 1.00 with an average mean factor congruence per solution (M_{mean}) of .998, and an average minimum factor congruence per solution (M_{min}) of .99. After varimax rotation, the factor congruence ranged overall between .09 and 1.00 ($M_{mean} = .999$, $M_{min} = .99$), and after promax rotation, it ranged overall between .10 and 1.00 ($M_{mean} = .99$, $M_{min} = .96$).

Despite the mostly small differences in loadings and pattern coefficients, the indicator-to-factor correspondences differed between the final promax-rotated R psych and SPSS solutions in 38.4% of all data sets and in 44.4% of the 187 data sets with a solution with two or more factors. This mostly concerned differences for 1 to 3 indicators, but it went up to differences for 50 indicators in the most extreme case. Finally, the mean and maximum absolute differences in pattern coefficients featured a credible association with the indicator-to-factor ratio in a Bayesian gamma regression

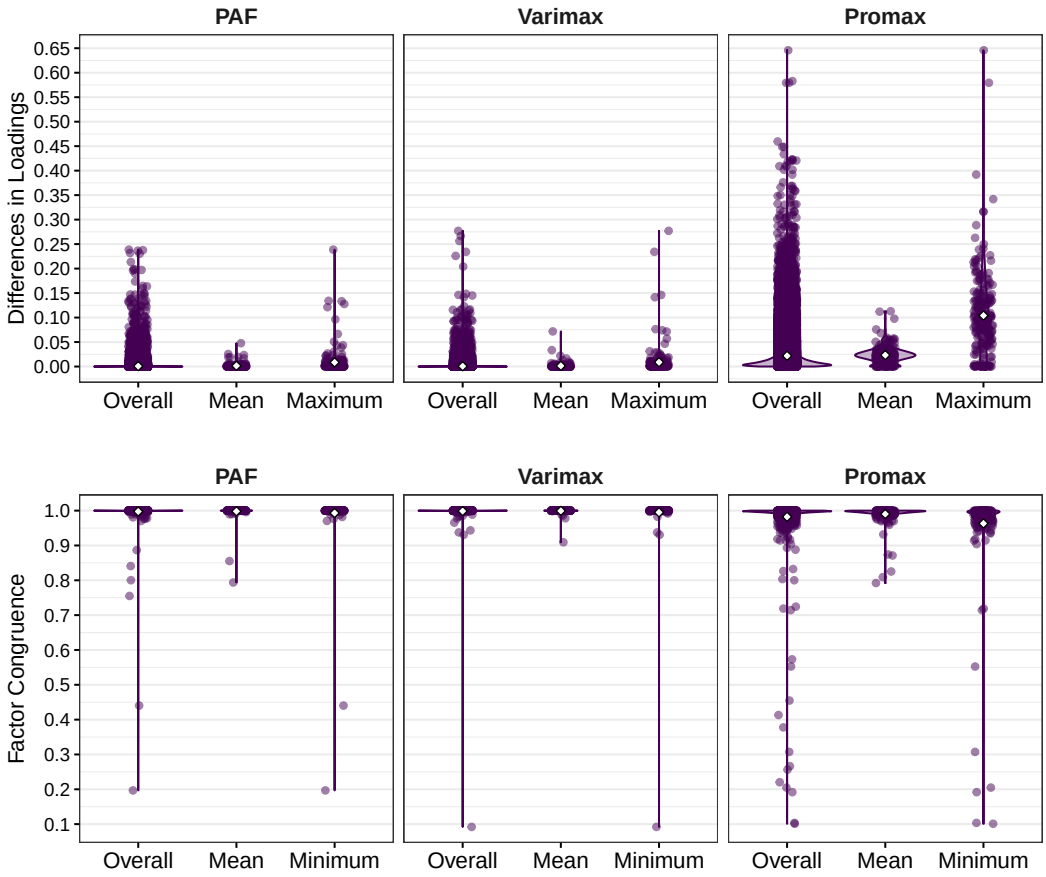

Figure 1. Overall, mean, and maximum differences in loadings/pattern coefficients per solution and overall, mean, and minimum factor congruence per solution from PAF without rotation, with varimax rotation, and with promax rotation, obtained with the R psych and SPSS implementations for 216 real data sets. The white diamond represents the mean. PAF = principal axis factoring.

analysis. A lower ratio was related to higher mean (b = -0.26, 95% HDI[-0.32; -0.21], r_s = -.40) and maximum (b = -0.23, 95% HDI[-0.30; -0.16], r_s = -.39) differences in pattern coefficients.

Discussion

When comparing the R psych and SPSS factor solutions for real data sets, we found mostly small differences in loadings and pattern coefficients and mostly large factor congruence. Still, for 38.4% of data sets the differences in pattern coefficients were large enough to result in different indicator-to-factor correspondences. Moreover, the number of factors for the first admissible solution in R psych and SPSS also differed in over a fifth of all data sets. The larger differences for pattern coefficients after promax rotation suggest that differences in results are mostly due to the different implementations of promax rotation in R psych and SPSS (i.e., using

an unnormalized versus normalized target matrix, respectively) and due to promax rotation amplifying pre-existing differences after varimax rotation, and less due to the different implementations of PAF and varimax rotation.

These results reveal that applying the (allegedly) same EFA procedure to the same data in the two programs can result in considerably different factor solutions. A researcher working with one program would thus often draw different conclusions concerning which indicator loads on which factor or even concerning the adequate number of factors to retain compared to a researcher working with the other program. In applications like scale construction, such differences might ultimately even lead to different scales.

It thus seems that the ways through the jingle jungle are indeed different enough to call for a guide towards the best way through. As these analyses were based on real data sets, we do not know the "true" population models behind them. It therefore remains unclear where these differences come from and whether there is an implementation that results in more accurate results in general. For example, certain properties of the data sets might influence differences in solutions between the implementations and the accuracy of results. The present results suggest that the indicator-to-factor ratio might be one such property. As a next step, we thus performed a simulation study comparing many different implementations of PAF and promax, including the R psych and SPSS implementation, to approach these questions systematically, and to find the best way through the jungle in the form of a recommendation about which implementation to use for most accurate results.

Differences in Accuracy between Implementations for Simulated Data

To be able to compare different implementations of PAF and promax in terms of their accuracy, a true model is needed for comparison. To this end, we simulated data based on different population models implementing various data structures. We then examined how well different implementations, featuring all possible combinations of the above identified settings of PAF and promax rotation, would recover the true solutions. That is, the aim of this simulation study was to search the space of possible implementations, including the R psych and SPSS ones, to test whether we could identify one implementation that would reliably yield more accurate solutions and would thus be preferable overall or at least for certain data structures. In addition, we also directly compared the R psych and SPSS implementations in more detail regarding their ability to recover the population models, as well as in terms of differences in pattern coefficients, in a separate simulation analysis. We report this latter analysis in the SM (section 3).

Methods

To compare the implementations regarding their ability to recover the underlying population model, we created 27 different pattern matrices and four different factor intercorrelation matrices, the combination of which resulted in 108 population models (see Table A2 for an overview, and SM, sections 5 and 6, for the pattern- and factor intercorrelation matrices). Therein, we varied (a) the number of factors,

(b) the number of indicators per factor, (c) the size of the pattern coefficients, (d) whether cross-loadings were present, and (e) the magnitude of the factor intercorrelations. Some of these population models were based on De Winter and Dodou (2012), yet we also added additional ones to cover a large space of possible data structures. For example, intelligence tests often exhibit a relatively low indicator-to-factor ratio and strong factor intercorrelations (e.g., Frazier & Youngstrom, 2007). Other measures, such as some personality scales, tend to have higher indicator-to-factor ratios with lower factor intercorrelations (e.g., Goldberg, 1999; Johnson, 2014). Simulating a diverse set of data structures permitted us to compare the implementations not only in general, but also on more specific levels, conditional on the data structure.

From each of these population models, we simulated two times 1,000 data sets from multivariate normal distributions with sample sizes of 180 and 450[6], respectively, for the model recovery. To this end, we used the following procedure to simulate each data set from a population model, which always consists of a combination of a population pattern matrix and a population factor intercorrelation matrix. We first obtained the population correlation matrix $\boldsymbol{\mathcal{R}}$ from the population pattern matrix $\boldsymbol{\Lambda}$ (see Table A2, and SM, section 5) and population factor intercorrelation matrix $\boldsymbol{\Phi}$ (see Table A2, and SM, section 6) with

$$\boldsymbol{\mathcal{R}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T \tag{1}$$

$$diag(\boldsymbol{\mathcal{R}}) = 1 \tag{2}$$

For simplicity, we set the variable means to zero and the SDs to 1, thus the population covariance matrix equals $\boldsymbol{\mathcal{R}}$. We then sampled data ($N = 180$ or $N = 450$) from a multivariate normal distribution based on the respective covariance matrix.

For each simulated data set, we conducted each of the different implementations of PAF with subsequent promax rotation by extracting the true number of factors of the population models (i.e., either three or six factors). Regarding the PAFs, we varied the following settings: Three initial communality estimates—unity, MAC, and SMCs; whether the absolute of the eigenvalues should be used or not; whether the convergence criterion is tested against the maximum difference in any single communality estimate, or against the difference in the sum of all communalities; and two different convergence criteria, namely 10^{-3} and 10^{-6}. Regarding the promax rotation, we varied the following settings: the two varimax types; whether the target matrix is normalized or not; and two different k parameters, namely (a) 4 in all cases versus (b) 3 in the case of unnormalized promax and 2 in the case of normalized promax (see Tataryn et al., 1999). This resulted in 192 possible implementations which we com-

[6]These sample sizes correspond to a ratio of sample size (N) to number of indicators p of 10 and 25 for our baseline model (18|3|6), respectively. According to MacCallum et al. (1999), a sample size of 180 is rather small for conditions like the ones in our baseline model. The smaller sample size and *N:p* ratio corresponds with the median values for our real data sets (see Table *Real_data_description.xlsx*, `https://osf.io/pcrqu/`). Conversely, a sample size of 450 should be sufficient also for more "problematic" data structures like those with low indicator-to-factor ratios and weak factors, as we simulated in some of our population models.

pared against each other in terms of their ability to recover the underlying population model.

Statistical Analyses. We compared the 192 different implementations in terms of the root mean squared errors (RMSE, i.e., the deviance of the fitted sample pattern matrix from the true population pattern matrix), the probability for the occurrence of Heywood cases, and the number of incorrect indicator-to-factor correspondences. The RMSE between the population pattern matrix $\mathbf{\Lambda}$ and a fitted sample pattern matrix $\hat{\mathbf{\Lambda}}$ was computed with

$$RMSE = \sqrt{\frac{trace[(\mathbf{\Lambda} - \hat{\mathbf{\Lambda}})^T(\mathbf{\Lambda} - \hat{\mathbf{\Lambda}})]}{pm}} \tag{3}$$

where p is the number of indicators, and m is the number of extracted factors.

We performed the following analyses separately for each of the 216 different models (108 population models times the two sample sizes). Moreover, for each of these models, we analyzed simulated data sets for which any of the m largest eigenvalues were negative during the PAF procedure for any of the implementations separately from those where all m largest eigenvalues were always positive.

We first ordered the different implementations from best to worst (i.e., lowest to highest), according to the average RMSE (MRMSE), proportion of Heywood cases, and number of incorrect indicator-to-factor correspondences, respectively. Next, we compared the best and the second best implementation with Bayesian regressions with a dummy coded identifier for the implementations as predictor. The regression model with the RMSE as dependent variable was implemented with a Gaussian family and an identity link function; the one with the probability of Heywood cases was implemented with a binomial family with a logistic link function; and the one concerning the number of incorrect indicator-to-factor correspondences was implemented with a negative binomial family with a log link function. All models were run using the *rstanarm* package (Goodrich et al., 2018) with default priors and results were inspected using the *bayestestR* package, version 0.7.2 (Makowski et al., 2019).

To test the credibility of effects, we employed the region of practical equivalence (ROPE) plus 95% HDI rule (Kruschke, 2018) for the linear and logistic regressions. That is, if the 95% HDI of a parameter fell completely outside the ROPE, we regarded this as conclusive evidence for an effect. If less than 95% of the HDI fell outside or inside the ROPE, we regarded this as inconclusive evidence for an effect. Finally, if the 95% HDI fell completely inside the ROPE, we regarded this as conclusive evidence that there was no effect. These analyses permitted us to judge whether one implementation was generally better suited for particular data structures. We defined the ROPE to be $[-0.1 * SD_{dv}, 0.1 * SD_{dv}]$ for linear regression models and $[-0.18, 0.18]$ for logistic regression models (in line with Kruschke, 2018; Makowski et al., 2019). Because there is no standard for choosing a ROPE for negative binomial regressions, we simply relied on the 95% HDI rule to gauge whether an effect was credible.

For analyses of data sets where all m largest eigenvalues were positive, regression analyses were only run if at least ten of the 1,000 simulated data sets per population

model resulted in all-positive m largest eigenvalues. Conversely, for analyses of data sets with some negative m largest eigenvalues, regression analyses were only run if at least ten of the 1,000 simulated data sets per population model resulted in some negative m largest eigenvalues. Moreover, a logistic regression was only run if additionally at least 1% of the data sets of the current population model contained a Heywood case, and a negative binomial regression was only run if additionally there were at least two unique numbers of incorrect indicator-to-factor correspondences present. This was to ensure that the models could be run and did not result in errors (e.g., due to no variance in the dependent variable).

If there was conclusive evidence for a difference between the best and second-best implementation, the procedure was stopped and counted as conclusive evidence for a difference between settings in the respective population model. If the evidence was inconclusive or there was conclusive evidence for equality of the solutions found by the two implementations, the best solution was compared to the third-best solution, and so on, either until conclusive evidence for a difference was found or until all 191 other implementations were compared against the best one. If after this procedure there was conclusive evidence for equality between the best and the worst implementation, we counted this as conclusive evidence for the absence of relevant differences between the implementations in the respective population model. If there was conclusive evidence neither for a difference nor for equality, the evidence was counted as inconclusive. Moreover, if no regression models could be run (e.g., because not a single Heywood case occurred for any implementation in a population model), differences were gauged descriptively. In these cases, we counted a complete match between all implementations (e.g., when none of the implementations resulted in a Heywood case) as conclusive evidence for equality, whereas an imperfect match was counted as inconclusive evidence.

To analyze the simulation results, we first descriptively compared the different implementations regarding their MRMSE, proportion of Heywood cases, and incorrect indicator-to-factor correspondences. Moreover, to determine which implementation on average produced the best results, we used the proportion across the 108 population models with which a given implementation was among the best ones (i.e., was the best implementation, had conclusive evidence for equality with the best implementation, or had inconclusive evidence for a difference to the best implementation) as determined with the regression analyses above.

Results

Results from our regression analyses revealed conclusive differences between at least some of the 192 implementations for the majority of the population models regarding RMSE and the probability of Heywood cases, and for about half of the population models regarding the correctness of indicator-to-factor correspondences (see Table 1). Although the differences in accuracies were mostly small, there were some population models—namely those with very low indicator-to-factor ratios as well as those with mixed or high factor intercorrelations—where differences were larger.

Moreover, there was no single implementation that was best across all con-

Table 1
Evidence for Differences in Accuracy, Proportion of Heywood Cases, and Correctness of Indicator-to-Factor Correspondences Between Implementations

	$N = 180$			$N = 450$		
Property	Inc	Eq	Diff	Inc	Eq	Diff
Data sets without negative eigenvalues						
RMSE	18	0	90	10	1	97
Heywood	20	22	66	13	44	51
Ind-to-Fac Corres	28	29	51	12	37	59
Data sets with negative eigenvalues						
RMSE	0	0	10	0	0	9
Heywood	0	0	10	0	0	9
Ind-to-Fac Corres	5	1	3	2	0	7

Note: Tally of type of evidence for the 216 different models (108 population models times 2 sample sizes) derived from Bayesian regression analyses. The row sum for the rows concerning data sets with negative eigenvalues are smaller than 216 because data sets with negative eigenvalues occurred only for some models. RMSE = Root mean squared error. Heywood = Probability of the occurrence of a Heywood case. Ind-to-Fac Corres = Difference in indicator-to-factor correspondence from found solution to population model. Inc = Inconclusive evidence. Eq = Conclusive evidence for no relevant difference between the implementations (equality). Diff = Conclusive evidence for a difference between at least some implementations.

ditions and regarding all three considered accuracy criteria. However, there were some settings that were clearly superior, namely sum for criterion type and a convergence criterion of 10^{-3} (both in the PAF procedure). A description and comparison of the implementations that produced the best results in one of the considered criteria is included in Table 2 and, for a larger set of implementations, in Table *best_implementations.xlsx* in the online repository, `https://osf.io/6prcz/`).

Table 2
Proportion of Population Models for Which the Two Best Implementations and the R psych and SPSS Implementations Were Among the Best

	Best	$\text{Best}_{\text{kaiser}}$	$\text{Psych}_{\text{unity}}$	$\text{Psych}_{\text{SMC}}$	SPSS
RMSE					
$N = 180$, pos. eigen.	.70	.70	.46	.58	.68
$N = 180$, neg. eigen.	.53	.53	.00	.00	.53
$N = 450$, pos. eigen.	**.72**	**.72**	.45	.59	.69
$N = 450$, neg. eigen.	.56	.56	.00	.00	.44
Heywood Cases					
$N = 180$, pos. eigen.	.56	.56	.69	.46	.53
$N = 180$, neg. eigen.	.07	.07	.33	.00	.07
$N = 450$, pos. eigen.	.63	.63	.81	.64	.62
$N = 450$, neg. eigen.	.44	.44	.67	.00	.44
Ind.-to-Fac. Corres.					
$N = 180$, pos. eigen.	.95	.95	.71	.83	.93
$N = 180$, neg. eigen.	**.64**	**.64**	.43	.00	.57
$N = 450$, pos. eigen.	.93	**.94**	.70	.76	.89
$N = 450$, neg. eigen.	**.89**	**.89**	.22	.00	**.89**
Settings					
PAF					
Communality method	SMC	SMC	unity	SMC	SMC
Criterion type	sum	sum	sum	sum	max. ind.
Absolute eigenvalues	yes	yes	no	no	yes
Convergence criterion	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
Promax rotation					
Varimax type	svd	kaiser	svd	svd	kaiser
P type	norm	norm	unnorm	unnorm	norm
k	4	4	4	4	4

Note: For positive eigenvalues, the proportion of the 108 population models for which the respective setting combination was among the best setting combinations is shown. For negative eigenvalues, the proportion of the population models including data sets that resulted in negative eigenvalues for which the respective setting combination was among the best setting combinations is shown. The top row contains the identifiers of the implementations, their settings are listed in the bottom part of the table. Boldface indicates that this implementation was most frequently among the best implementations for the respective data sets. $\text{Best}/\text{Best}_{\text{kaiser}}$ = implementations with best results overall, with varimax type svd and kaiser, respectively; $\text{Psych}_{\text{unity}}/\text{Psych}_{\text{SMC}}$ = R psych implementation with unity/SMC as initial communality estimates; SPSS = SPSS implementation; RMSE = root mean square error; pos. eigen. = all-positive eigenvalues; neg. eigen. = some negative eigenvalues; Ind.-to-Fac. Corres. = indicator-to-factor correspondences; PAF = principal axis factoring; P type = target matrix type; k = power in promax; MAC = maximum absolute correlation; SMC = squared multiple correlation; sum = deviance of the sum of all communalities; max. ind. = maximum absolute deviance of any communality; unnorm = unnormalized; norm = normalized; svd = singular value decomposition.

Regarding RMSE, the implementations that produced the most accurate results differed for the sample sizes of 180 and 450. For PAF, the best implementations in the case of $N = 180$ all used MACs as initial communality estimates, sum as criterion type, and a convergence criterion of 10^{-3} (both treatments of eigenvalues produced equally accurate results). For the promax rotation, they used an unnormalized target matrix with $k = 3$ (both varimax types produced equally accurate results). When

the sample size was 450, SMCs were used as initial communality estimates instead of MACs, and only absolute eigenvalues were used—the other settings were the same. For the promax rotation, the best implementations then used a normalized target matrix with $k = 4$ (again, both varimax types produced equally accurate results).

Regarding Heywood cases, unity as initial communality estimates produced the best results. The other PAF settings were the same as the best settings regarding RMSE. Moreover, using SMCs as initial communality estimates tended to result in more Heywood cases compared to using MACs. For promax, a normalized target matrix with $k = 2$ produced the lowest proportion of Heywood cases—again irrespective of the varimax type.

Finally, regarding indicator-to-factor correspondences, the best implementations for PAF were again ones that used sum as criterion type and a convergence criterion of 10^{-3} as initial communality estimates. Together with MACs, it did not matter whether absolute eigenvalues were used or not. Together with SMCs, it was clearly better to use absolute eigenvalues. For promax, a normalized target matrix with $k = 4$ was best—also irrespective of the varimax type used.

On average, the impact of the different implementations, although in many cases statistically robust (see Table 1), tended to be smaller than that of the data structures—that is, the population models (see Figure 2 and Figure S6). To gauge the impact of the data structures, we computed the differences between the best- and worst-performing implementation in a given population model, separately for every population model.[7] Conversely, to gauge the impact of the implementations, we computed the differences between the population models leading to the best and worst performance in a given implementation, separately for each implementation. In this analysis we focused on the data sets with $N = 450$.

Based on this approach, we found that across all implementations the differences in MRMSE ranged from .00 to .09 ($Mdn = .01$; $M = .02$). In contrast, across the population models, the differences in MRMSE ranged from .18 to .28 ($Mdn = .21$; $M = .22$). Moreover, the difference in the proportion of Heywood cases across all implementations ranged from .00 to .92 ($Mdn = .01$; $M = .17$), and the difference in proportion of Heywood cases across the population models ranged from .28 to .98 ($Mdn = .87$; $M = .80$; see also Figure 3 and Figure S7). Finally, across the implementations the differences in the proportion of solutions with at least one incorrect indicator-to-factor correspondence ranged from .00 to .50 ($Mdn = .01$; $M = .05$), while, across population models, this difference in proportions was always 1.00 (see also Figure 4 and Figures S8–S10).

Overall, the different implementations resulted in solutions with larger discrepancies for population models with high factor intercorrelations, with low pattern coefficients, with cross-loadings, and with low indicator-to-factor ratios; that is, exactly the data structures that also led to larger errors and to higher proportions of Heywood cases. In the other data structures, the implementations converged more

[7]Note that therefore no single implementation was used as best for every population model, but the respective implementation that in the current population model, across the 1,000 simulated data sets, resulted in the lowest MRMSE, proportion of Heywood cases, or proportion of solutions with at least one incorrect indicator-to-factor correspondence.

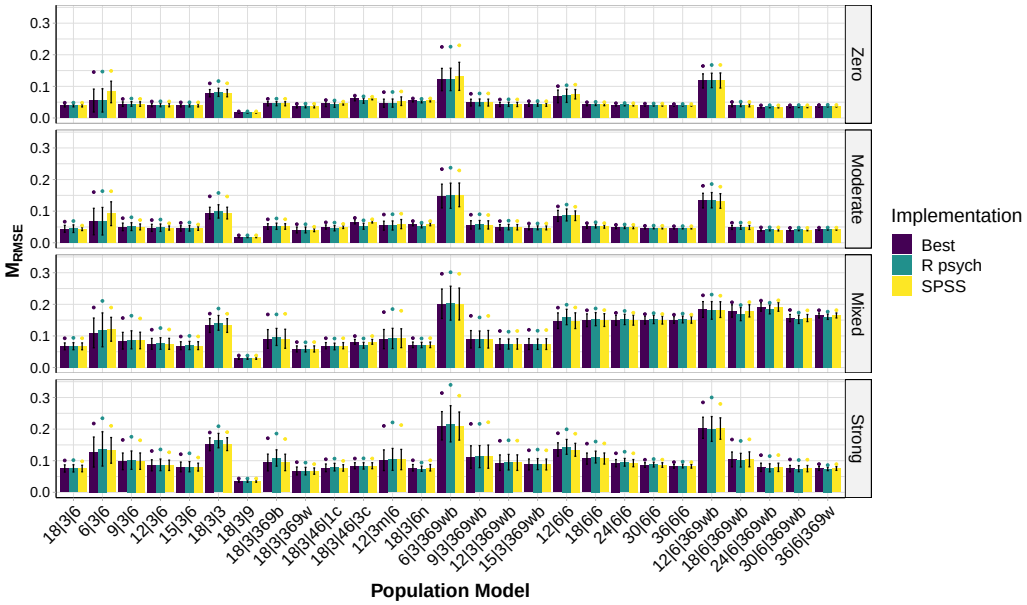

Figure 2. Distributions of RMSE of three different implementations, separately for the different population models, based on data sets simulated with $N = 450$. A population model is always a combination of a pattern matrix (on the x-axis) and a factor intercorrelation matrix (the facets). Bars indicate the mean RMSE, the whiskers indicate $\pm 1SD$, and dots represent the mean of the 5% largest RMSE. Best = implementation with best results overall; R psych = R psych implementation with SMC as initial communality estimates when no negative eigenvalues occurred and R psych implementation with unity as initial communality estimates when negative eigenvalues occurred; SPSS = SPSS implementation. See Table 2 for more information on these implementations. An overview of the population models is provided in Table A2. The detailed population pattern- and factor intercorrelation matrices are provided in the SM, sections 5 and 6.

strongly.

Discussion

Our simulations to compare many possible implementations of PAF and promax led to two main insights. First, we found that many statistically robust differences between implementations occurred and were able to pinpoint properties of implementations that led to the best results regarding RMSE, Heywood cases, and indicator-to-factor correspondences. The best trade-off seems to be to use SMCs as initial communality estimates, sum as criterion type, absolute eigenvalues, and a convergence criterion of 10^{-3} for PAF; and a normalized target matrix with $k = 4$ for promax with any of the two varimax types (*best* and $best_{kaiser}$ in Table 2). One of these implementations—the one named *best* in Table 2—is implemented as default in the *EFAtools* R package (Steiner & Grieder, 2020), along with the possibility of varying the settings as done in this simulation study.

With smaller samples, using MACs instead of SMCs and an unnormalized tar-

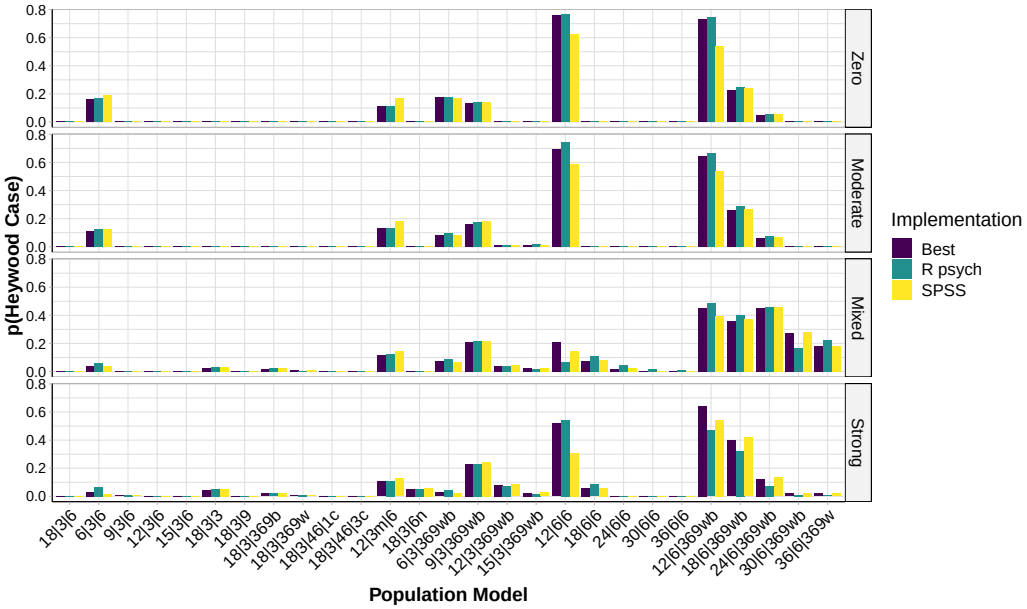

Figure 3. Proportions of solutions per implementation out of the 1000 simulated data sets in which Heywood cases occurred, separately for the different population models, based on data sets simulated with $N = 450$. A population model is always a combination of a pattern matrix (on the x-axis) and a factor intercorrelation matrix (the facets).Best = implementation with best results overall; R psych = R psych implementation with SMC as initial communality estimates when no negative eigenvalues occurred and R psych implementation with unity as initial communality estimates when negative eigenvalues occurred; SPSS = SPSS implementation. See Table 2 for more information on these implementations. the population models is provided in Table A2. The detailed population pattern- and factor intercorrelation matrices are provided in the SM, sections 5 and 6.

get matrix with $k = 3$ could be beneficial to maximize the accuracy of the pattern coefficients. However, this comes at the cost of a higher probability for erroneous indicator-to-factor correspondences. If the aim is to minimize the chance of Heywood cases, one could try with the above-mentioned implementation first and if this implementation leads to Heywood cases, try again with MACs as initial communalities and if this still leads to Heywood cases, try again with unity and with $k = 2$. However, although the latter will minimize the chance for the occurrence of Heywood cases, it will likely come with the cost of less accurate pattern coefficients and a larger number of erroneous indicator-to-factor correspondences.

The second main insight is that the data structures had a large impact on the recovered factor solutions. For example, a low indicator-to-factor ratio often led to problems such as worse fit, the occurrence of Heywood cases, and problems in the identification of the correct indicator-to-factor correspondences. Moreover, solutions with weak factors also displayed worse fit. Finally, large factor intercorrelations and weak factors led to severe problems in recovering the correct indicator-to-factor correspondences, which might limit the robustness of EFA (or at least of PAF and promax

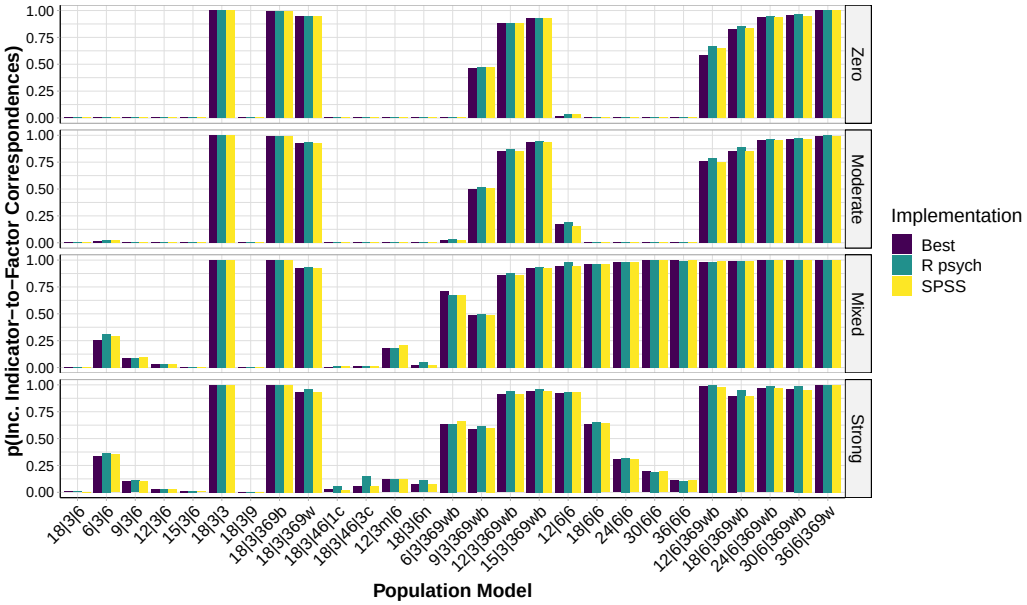

Figure 4. Proportions of solutions per implementation out of the 1'000 simulated data sets with at least one incorrect indicator-to-factor correspondence, separately for the different population models, based on data sets simulated with $N = 450$. A population model is always a combination of a pattern matrix (on the x-axis) and a factor intercorrelation matrix (the facets). Best = implementation with best results overall; R psych = R psych implementation with SMC as initial communality estimates when no negative eigenvalues occurred and R psych implementation with unity as initial communality estimates when negative eigenvalues occurred; SPSS = SPSS implementation. See Table 2 for more information on these implementations. An overview of the population models is provided in Table A2. The detailed population pattern- and factor intercorrelation matrices are provided in the SM, sections 5 and 6.

rotation) for this kind of data structures.

To sum up, we succeeded in finding a best way through the jingle jungle in the form of the best implementation identified here. Another approach might be to incorporate the full jungle into one's analyses by using model averaging. We further discuss this possibility below.

General Discussion

We compared an EFA procedure with PAF and promax rotation between the two most prominently used programs in psychological research (Dunn, 2011): R (using the popular *psych* package) and SPSS. We indeed discovered a jingle jungle for the investigated EFA procedure and the programs considered: Equal names do not mean equal implementations, and with this do not necessarily mean comparable results. But how different are the results? And is there a best way through the jungle?

Our main findings can be summarized in three points. First, we found many systematic differences in pattern coefficients and in accuracy between the tested im-

plementations, including the R psych and SPSS implementations. Although most differences in (the accuracy of) pattern coefficients were small, very large ones occurred for some data sets and population models, and many led to differences in (the accuracy of) indicator-to-factor correspondences. Second, neither of the two implementations—R psych or SPSS—consistently resulted in more accurate solutions than the other across all population models. The implementations producing the most accurate results for PAF and promax rotation are combinations of the R psych and SPSS implementations. Third, the data structure is at least as important as the implementation for the accuracy of results. For some data structures, the accuracy is very low, regardless of the implementation.

Mostly Small, But Systematic Differences

As is evident from results from both the real data analysis and the simulation study, the different implementations—including R psych and SPSS—will lead to comparable results in many cases. Nevertheless, we found pattern coefficients differed systematically between implementations, with differences in accuracy for 88% of the population models. Moreover, for the real data sets, the average maximum difference between the R psych and SPSS solutions was a non-negligible 0.10. Perhaps even more importantly, the indicator-to-factor correspondences differed between R psych and SPSS for 38% of the real data sets (44% of those with at least two factors), and their accuracy differed between implementations for 51% of the population models in our simulation study. Due to the use of thresholds, even small differences in pattern coefficients might affect indicator-to-factor correspondences and with this even decisions on which solution to retain. A strict use of thresholds is of course questionable. Nevertheless, thresholds are applied and they may be used to too strongly defend a more or less arbitrary choice about which solution to retain, especially if this solution is in accordance with a favored theory.

Overall, our results demonstrate that different implementations of PAF and promax rotation—with promax rotation likely having the greater impact—can lead to fairly different solutions, and in many cases also to different conclusions regarding the factor structure of the investigated construct. Given the amount and impact of these differences, a natural question to ask is whether there is one implementation of PAF and promax rotation that renders the most accurate results under different conditions.

Best PAF and Promax Implementations

As stated above, none of the considered implementations—including R psych and SPSS—consistently outperformed all others. Yet, our simulation analyses still permitted us to identify specific settings that seem to be advantageous and allowed us to make a general recommendation for implementations of PAF and promax rotation. These implementations constitute a combination of the R psych and SPSS implementations. Specifically, for PAF, one should take the absolute eigenvalues (as SPSS does), compute initial communalities with SMCs, use 10^{-3} as convergence criterion, and apply the convergence criterion to the sum of all communalities (as R

psych does). For promax rotation, a row-normalization should be done on the target matrix of the varimax solution (row-normalized promax rotation, as done in SPSS) and $k = 4$ should be used as the power to which to raise the elements of the target matrix. These combinations of settings are implemented as default for PAF and promax rotation in the *EFAtools* package.

Although a general recommendation for these implementations is justified, our results also reveal that the choice of one implementation over the other will probably have a major impact on results only in certain cases. Our recommended implementation will therefore maximize the verisimilitude probably especially for more "problematic" data structures like the ones discussed in the next paragraph.

Data Structure Is At Least As Important As Implementation

Our simulation analyses revealed that the data structure had a strong influence on the accuracy of a solution. For cases with a low indicator-to-factor ratio, weak factors, and large factor intercorrelations (independent of each other), the accuracy of results was often very limited, to the point that, for some cases, all of the 1,000 solutions had at least one incorrect indicator-to-factor correspondence (see Figure 4).

The finding that a low indicator-to-factor ratio can be problematic in EFA is in line with previous research (e.g., MacCallum et al., 1999). Specifically, a low indicator-to-factor ratio results in less stable and less replicable factor solutions (Gorsuch, 1983; MacCallum et al., 1999; Mulaik, 2010; Tucker & MacCallum, 1997). Regarding factor intercorrelations, it has been shown that a positive manifold (i.e., exclusively positive intercorrelations) leads to better recovery of factor solutions (Tucker & MacCallum, 1997). Our results somewhat contradict these findings: In our simulation study, factor structures with mixed (.30, .50, .70) and with high (.70) factor intercorrelations resulted in a less accurate recovery of the true factor solutions, while orthogonal factors caused the least problems. This is true for both investigated sample sizes. However, others have also proclaimed (Gorsuch, 1983) and shown (Gerbing & Hamilton, 1996) that too high factor intercorrelations can be problematic in factor analysis, especially for the recovery of weak factor loadings (De Winter & Dodou, 2012). These findings are corroborated by our analyses, where factors with low pattern coefficients (.30) were also worse recovered compared to factors with higher pattern coefficients (.60 or .90), especially if factor intercorrelations were high. Previous studies focusing exclusively on orthogonal factor structures have also demonstrated worse recovery of weak compared to stronger factors (Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005; MacCallum et al., 1999). Finally, the data structure also influences the sample size necessary for a stable factor solution. That is, a stable solution can be achieved with smaller samples if the indicator-to-factor ratio is high, if communalities are high (strong factors), and if the factors are correlated (Hogarty et al., 2005; MacCallum et al., 1999; Tucker & MacCallum, 1997).

Despite these findings, problematic data structures, such as a low indicator-to-factor ratio, became increasingly common in factor-analytic research, especially on intelligence tests (Frazier & Youngstrom, 2007) and are still quite common to date (Goretzko, Pham, & Bühner, 2019). It is therefore important that simulation studies

in a factor-analytic framework take these problematic data structures into account.

Model Averaging: An Alternative Way Through the Jungle?

Given that the variation in factor solutions across implementations was larger for problematic data structures, the amount of variability between different implementations might be useful to judge the stability of the factor solution for a particular data set. Thus, applying multiple implementations of the same procedure, and possibly also different extraction and rotation methods[8], could serve as a robustness check for a factor structure. A larger variation in the factor solutions across the different implementations or methods would indicate a more unstable factor structure for this particular data set and could render it more likely that the data structure is problematic. Researchers could then use this as a guide to judge whether more indicators or more participants should be sampled.

Applying multiple implementations or methods with a common purpose also allows to calculate an average factor solution across all different implementations and methods, which could be preferable to results from a single implementation as it takes model uncertainty into account. With this, we get in the realm of model averaging, which Fletcher (2018) defines as "a means of allowing for model uncertainty in estimation which can provide better estimates and more reliable confidence intervals than model selection" (p. 1). It typically involves calculating a weighted average of the model parameters, with stacking as the preferred method of weighting for frequentist model averaging, and another common method being AIC weights (Fletcher, 2018). Model averaging has mainly been used in regression frameworks and on models with different sets of predictor variables (Fletcher, 2018; Schomaker & Heumann, 2011; Steel, 2020). For EFA, there are some studies using averaging to get more robust estimates (e.g., Gerbing & Hamilton, 1996 who averaged across different rotation methods), and one study used model averaging with AIC weights across ML estimated solutions with different numbers of factors (Schomaker & Heumann, 2011). To our knowledge, however, there is no study that investigated model averaging across different settings or methods within the EFA framework.

In sum, model averaging might be a possibility not only to provide a straight path through the jingle jungle ignoring the surroundings, like our identified best implementation, but rather to produce a full map of the jungle, and possibly also an even better way out of it in the form of an average factor solution. For an easy application and first step towards model averaging we include a function in the *EFAtools* package to flexibly perform and average across different implementations of PAF, varimax, and promax rotation, as well as different extraction and rotation methods with unit-weighting. Clearly, future research is needed to investigate how model averaging is best implemented for EFA to maybe result in more stable and reliable parameter estimates compared to using a single EFA model. One challenge in this will be to identify which weights to use for averaging. For PAF, AICs cannot be used sensibly, and stacking is not practicable for EFA, either. Until these questions are answered,

[8]For rotation methods this would, of course, only make sense for the same type of rotation; that is, either multiple orthogonal methods or multiple oblique methods.

this kind of model averaging may mainly be useful for testing the robustness of a factor solutions across different implementations, and researchers may want to rely on the best implementation identified here as a best guess for the model parameters.

Implications

One implication of our study is that researchers should not assume that procedures with the same name are implemented the same way and necessarily lead to comparable results. A consequence of such incomparable results from factor-analytical methods is that different researchers might draw different conclusions about the latent structure of their data, depending on which program they used. This could be especially problematic if EFA is used to create new instruments or as a tool in theory building. Moreover, it could lead to results appearing not to be replicable as a consequence and as such might even contribute to the replication crisis (see Ershova & Schneider, 2018; Flake et al., 2017). If different programs are thought to lead to comparable results, differences in implementations might not be considered as an explanation for the failure to replicate an analysis. This bears the potential for waste of money and time to find conceptual or methodological explanations for differences that might actually just be an artifact of the procedure used.

To counteract possible misconceptions and to facilitate comparisons of implementations of the kind performed here, we advocate the use of free and open-source software (e.g., R, python, or julia) and the sharing of analysis scripts. Doing so empowers other researchers to track the analyses and, if questions arise, to dig into the code of the actual procedures. Such comparisons are harder to do with proprietary software as one needs to have the appropriate license and even then one has to rely on reconstructions of the procedures in most cases because the source-code is not publicly available. Another advantage of the use of open-source software is that newly developed or enhanced procedures can be shared immediately.

Our results also demonstrate that researchers should be cautious when interpreting their EFA results. They might be overconfident in their factor-analytical results and what these tell them about the "real" structure of the assumed underlying latent construct (see Yarkoni, 2019). However, especially for data structures like those often present in intelligence tests (low indicator-to-factor ratio, high factor intercorrelations, and weak first-order factors), EFA seems to have difficulties recovering the data structure, even when there exists a "true" underlying model and data are simulated from multivariate normal distributions, as in our simulations. Future studies could vary more properties of both real and simulated data sets. For example, it would be interesting to see how the presence of ordinal data or departures from (multivariate) normality, in combination with different correlation methods, might influence results.

Given these issues, researchers might be tempted to attribute these to the exploratory nature of EFA and to instead put their trust in (presumably) less data-driven methods, such as confirmatory factor analysis (CFA). However, the same kind of data structure that poses difficulties to EFA is problematic for CFA as well (e.g., Marsh, Hau, Balla, & Grayson, 1998; Ximénez, 2006, 2009). This is not surprising,

given that EFA and CFA are not that different after all and neither is necessarily more data- or more theory-driven than the other (Schmitt, 2011).

Researchers intending to use factor-analytic methods should therefore design their tests and questionnaires such that the data are suitable for factor analysis; that is, ensure a large enough indicator-to-factor ratio (probably at least 5:1 or 6:1; e.g., Goretzko et al., 2019; Gorsuch, 1983) and a sufficiently large sample size (at least 400) that should be larger the smaller the indicator-to-factor ratio is, and if weak factors are expected, especially when paired with high expected factor intercorrelations (De Winter & Dodou, 2012; Goretzko et al., 2019; Gorsuch, 1983; Hogarty et al., 2005; MacCallum et al., 1999; Tucker & MacCallum, 1997). Thus, ensuring an appropriate data structure is a prerequisite for a stable and replicable factor solution. Last but not least, a factor structure identified with factor-analytic methods should always be tested against external criteria to ensure its validity.

Limitations

One limitation of our work is that we only investigated a specific procedure within one statistical framework—one extraction method and one rotation method out of many, even though they are among the most prominent ones within EFA. But even within this narrow selection of procedures we found several differences in the implementations and the resulting solutions. Future research could investigate to what extent these issues apply to other factor-analytical methods (e.g., ML estimation) and statistical procedures. Similar jingle jungles have already been found between R and SPSS, for example, for linear regression (krissen, 2018; u/kriesniem, 2018), cox regression (GaryStats, 2017), multinomial logistic regression (Collins, 2016), and ANOVA (del Rio, 2017). Our guess is that many further jingle jungles might be found for other programs and statistical procedures. In fact, similar jungles have already been discovered for nonparametric statistical procedures across the four programs SPSS, SAS, Stata, and R, where the authors also point to the consequences of such variations for replication attempts (Hodges et al., 2020).

We could also only compare a selection of the various programs available to perform these analyses. As we chose the programs that are probably the most used in psychological research (Dunn, 2011), and as the psych package has also influenced the implementation of the EFA framework in python (Biggs, 2019), our findings are likely relevant for many researchers performing EFA. Moreover, we also went beyond the implementations of these two programs and included all possible combinations of the identified settings for PAF and promax in our simulation analysis.

Another limitation is that we did not have access to the source code of SPSS and were therefore only able to reproduce its implementations by implementing mathematical formulas from the algorithms manual (IBM Corp., 2017) and comparing the original output from SPSS with the one from our reconstruction of the SPSS implementation. Especially for the varimax implementation, it was difficult to find out how exactly the procedure is implemented in SPSS, as results with the implementation based on the formulas in the algorithms manual were not comparable enough to the original SPSS results. After some adjustments to the varimax criterion, we managed

a closer reproduction of original SPSS results. However, without access to the source code, it was impossible for us to determine the exact deviations of the implementation from the formulas provided in the algorithms manual. Nevertheless, we were able to reproduce results from SPSS with high accuracy and therefore believe that we reconstructed its algorithms well enough for the purpose of our analyses.

Finally, as is always the case for simulation studies, it is unclear how results from our simulation analyses generalize to other data structures and sample sizes not simulated here. However, we varied many different characteristics of the data structures which reflect many data structures occurring in EFA research (e.g., Goretzko et al., 2019). Similarly, it is unclear to what extent results from simulation studies generalize to real data analyses, where there is no true model underlying the data, "everything is correlated with everything" (Meehl, 1990, , p. 123), and thus an infinitely large number of constructs may influence a correlational structure (Cudeck & Henly, 1991; MacCallum, 2003; Meehl, 1990; Preacher, Zhang, Kim, & Mels, 2013). This caveat is one reason why we also analyzed real data sets. The agreement of the results from our real data analyses, where we included a large number of diverse data sets, and from the simulation study might at least indicate some generalizability of results from our simulation study to real-world problems.

Conclusion

Our results show that a jingle fallacy is indeed apparent in the investigated EFA procedure. That is, EFA methods named the same are actually implemented differently in the programs considered. The jingle jungle we discovered does have important implications, with different implementations frequently leading to different conclusions regarding the factor structure. We advocate the search for and exploration of further jingle jungles to gauge the extent of this issue for other statistical procedures and to enable researchers to make an informed decision about which program or implementation to use best. Moreover, we encourage researchers to state which version of a program they used for a particular analysis—as details in the implementations might be subject to change over time—as well as to familiarize themselves with default values provided in software packages to understand which specifications are used in their analyses. With the present work, we hope to raise awareness of possible differences in implementations of statistical procedures in different programs, and we advise researchers to use our recommended settings for PAF and promax rotation as the best way through the jungle, at least until other, potentially better ways through it—such as model averaging—have been tested sufficiently.

References

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. doi: 10.1037/met0000200

Biggs, J. (2019). *factor_analyzer: A python module to perform exploratory factor analysis* (Version 0.3.1). Princeton, NJ: Educational Testing Service. Retrieved from `https://factor-analyzer.readthedocs.io/en/latest/index.html`

Braeken, J., & Van Assen, M. A. (2017). An empirical kaiser criterion. *Psychological Methods*, *22*(3), 450. doi: 10.1037/met0000074

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, *38*, 25–56. doi: 10.1207/S15327906MBR3801_2

Burt, C. (1948). The factorial study of temperamental traits. *Journal of Statistical Psychology*, *1*, 178–203. doi: 10.1111/j.2044-8317.1948.tb00236.x

Carroll, J. B. (1958). Solution of the oblimin criterion for oblique rotation in factor analysis. *Unpublished manuscript*.

Collins, J. (2016). *Multinomial logistic regression in r vs spss.* Retrieved from `https://stats.stackexchange.com/questions/189424/multinomial-logistic-regression-in-r-vs-spss`

Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, *70*(6), 885–901. doi: 10.1177/0013164410379332

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*(3), 512–519. doi: 10.1037/0033-2909.109.3.512

Cureton, E. E. (1976). Studies of the Promax and Optres rotations. *Multivariate Behavioral Research*, *11*, 449–460. doi: 10.1207/S15327906MBR3403_3

del Rio, E. (2017). *Comparison of R and SPSS: ANOVA.* Retrieved from `https://medium.com/humansystemsdata/analysis-of-variance-showdown-r-vs-spss-f4e50234a94`

De Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*, 695–710. doi: 10.1080/02664763.2011.610445

Dinno, A. (2018). paran: Horn's test of principal components/factors (Version 1.5.2) [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=paran`

Dunn, T. (2011). The use of 'R' statistical software in psychology research. *PsyPAG Quarterly*, *81*, 10–13. Retrieved from `https://bgro.collections.crest.ac.uk/204/6/Dunn_UseOfR_2011.pdf`

Dwyer, P. S. (1939). The contribution of an orthogonal multiple factor solution to multiple correlation. *Psychometrika*, *4*, 163–171. doi: 10.1007/BF02288494

Ershova, A., & Schneider, G. (2018). Software updates: The "unknown unknown" of the replication crisis. *Impact of Social Sciences Blog*. Retrieved from `eprints.lse.ac.uk/90750/1/Ershova_Software-updates_Author.pdf`

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*,

4, 272–299. doi: 10.1037/1082-989X.4.3.272

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. doi: 10.1177/1948550617693063

Fletcher, D. (2018). *Model averaging*. Berlin, Germany: Springer.

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence*, *35*(2), 169–182. doi: 10.1016/j.intell.2006.07.002

GaryStats. (2017). *Why are SPSS and R producing different results for a cox regression on the same data, with the same model specification?* Retrieved from `https://stats.stackexchange.com/questions/263425/why-are-spss-and-r-producing-different-results-for-a-cox-regression-on-the-same`

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 62–72. doi: 10.1080/10705519609540030

Gernsbacher, M. A. (2018). Rewarding research transparency. *Trends in Cognitive Sciences*, *22*, 953–956. doi: 10.1016/j.tics.2018.07.002

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). *rstanarm: Bayesian applied regression modeling via Stan* (Version 2.21.1).

Goretzko, D., Pham, T. T. H., & Bühner, M. (2019). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, 1–12. doi: 10.1007/s12144-019-00300-2

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.

Gosh, A. (2019). *What's the best statistical software? a comparison of r, python, sas, spss and stata.* Retrieved from `https://www.inwt-statistics.com/read-blog/comparison-of-r-python-sas-spss-and-stata.html`

Grieder, S., & Grob, A. (2020). Exploratory factor analysis of the Intelligence and Development Scales–2: Implications for theory and practice. *Assessment*, *28*(8), 1853–1869. doi: 10.1177/1073191119845051

Guttman, L. (1956). "Best possible" systematic estimates of communalities. *Psychometrika*, *21*, 273–285. doi: 10.1007/BF02289137

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago, IL: University of Chicago Press.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205. doi: 10.1177/1094428104263675

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*, 65–70. doi: 10.1111/j.2044-8317.1964.tb00244.x

Hodges, C. B., Stone, B. M., Johnson, P. K., Carter, J. H., & Lindsey, H. M. (2020). *Researcher degrees of freedom and a lack of transparency contribute to unreliable results of nonparametric statistical analyses across SPSS, SAS, Stata, and R.* PsyArXiv. doi: 10.31234/osf.io/zem2w

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, *65*(2), 202–226. doi: 10.1177/0013164404267287

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. doi: 10.1007/BF02289447

Huang, F. (2015). hornpa: Horn's (1965) test to determine the number of components/factors (Version 1.0) [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=hornpa`

IBM Corp. (2014). IBM SPSS Statistics 23 algorithms [Computer software manual]. Armonk, NY.

IBM Corp. (2016). IBM SPSS Statistics 24 algorithms [Computer software manual]. Armonk, NY.

IBM Corp. (2017). IBM SPSS Statistics algorithms [Computer software manual]. Armonk, NY.

IBM Corp. (2020). *IBM SPSS Statistics for Macintosh, Version 27.0* [Computer software]. Armonk, NY.

IBM Support. (2020). *Can SPSS Statistics produce McDonald's omega reliability coefficient?* Retrieved from `https://www.ibm.com/support/pages/can-spss-statistics-produce-mcdonalds-omega-reliability-coefficient`

Jennrich, R. I. (2001). A simple general procedure for orthogonal rotation. *Psychometrika*, *66*(2), 289–306. doi: 10.1007/BF02294840

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, *71*, 173–191. doi: 10.1007/s11336-003-1136-B

Jennrich, R. I., & Sampson, P. (1966). Rotation for simple loadings. *Psychometrika*, *31*, 313–323. doi: 10.1007/BF02289465

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78–89. doi: 10.1016/j.jrp.2014.05.003

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200. doi: 10.1007/BF02289233

Kline, P. (1997). *An easy guide to factor analysis.* London, England: Routledge.

Klinke, A., Sigbertand Mihoci, & Härdle, W. (2010). Exploratory factor analysis in MPlus, R, and SPSS. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute. Retrieved from `https://www.researchgate.net/profile/Sigbert_Klinke/publication/228426652_EXPLORATORY_FACTOR_ANALYSIS_IN_MPLUS_R_AND_SPSS/links/0deec51ac615893693000000.pdf`

Kotenko, I. (2017). Same statistical method different results? Don't panic the reason might be obvious. Retrieved from `https://www.pharmasug.org/proceedings/2017/QT/PharmaSUG-2017-QT12.pdf`

krissen. (2018). *Difference between R and SPSS linear model results.* Retrieved from `https://stackoverflow.com/questions/53868465/difference-between-r-and-spss-linear-model-results`

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation.

Advances in Methods and Practices in Psychological Science, *1*(2), 270–280. doi: 10.1177/2515245918771304

Lorenzo-Seva, U. (1999). Promin: A method for oblique factor rotation. *Multivariate Behavioral Research*, *34*, 347–365. doi: 10.1207/S15327906MBR3403_3

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*(2), 57–64. doi: 10.1027/1614-1881.2.2.57

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364. doi: 10.1080/00273171.2011.564527

MacCallum, R. C. (1983). A comparison of factor analysis programs in SPSS, BMDP, and SAS. *Psychometrika*, *48*, 223–231. doi: 10.1007/BF02294017

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139. doi: 10.1207/S15327906MBR3801_5

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99. doi: 10.1037/1082-989X.4.1.84

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. doi: 10.21105/joss.01541

Marr-Lyon, L. R., Gupchup, G. V., & Anderson, J. R. (2012). An evaluation of the psychometric properties of the Purdue Pharmacist Directive Guidance Scale using SPSS and R software packages. *Research in Social and Administrative Pharmacy*, *8*, 166–171. doi: 10.1016/j.sapharm.2011.01.001

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? the number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*(2), 181–220. doi: 10.1207/s15327906mbr3302_1

Meehl, P. E. (1990). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. doi: 10.1207/s15327965pli0102_1

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. doi: 10.1038/s41562-016-0021

Newsom, J. (2020). *Exploratory factor analysis example.* Retrieved from `http://web.pdx.edu/~newsomj/semclass/ho_efa%20example.pdf`

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi: 10.1126/science.aac4716

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*(3), 396–402. doi: 10.3758/BF03200807

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28–56. doi: 10.1080/00273171.2012.710386

R Core Team. (2020). *R: A language and environment for statistical computing, Version 4.0.3* [Computer software]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raiche, G., & Magis, D. (2020). *nfactors: Parallel analysis and other non graphical solutions*

to the Cattell scree test (Version 2.4.1). Retrieved from `https://CRAN.R-project.org/package=nFactors`

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.0.12). Evanston, IL. Retrieved from `https://CRAN.R-project.org/package=psych`

Roff, M. (1936). Some properties of the communality in multiple factor theory. *Psychometrika*, *1*, 1–6. doi: 10.1007/BF02287999

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61. doi: 10.1007/BF02289209

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304–321. doi: 10.1177/0734282911406653

Schomaker, M., & Heumann, C. (2011). Model averaging in factor analysis: an analysis of olympic decathlon data. *Journal of Quantitative Analysis in Sports*, *7*(1). doi: 10.2202/1559-0410.1249

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*, 201–292. doi: 10.2307/1412107

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Stanley, D. (2015). *Ensuring R generates the same ANOVA F-values as SPSS.* Retrieved from `http://www.statscanbefun.com/rblog/2015/8/27/ensuring-r-generates-the-same-anova-f-values-as-spss`

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, *58*(3), 644–719. doi: 10.1257/jel.20191385

Steiner, M. D., & Grieder, S. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53), 2521. doi: 10.21105/joss.02521

Tataryn, D. J., Wood, J. M., & Gorsuch, R. L. (1999). Setting the value of k in promax: A Monte Carlo study. *Educational and Psychological Measurement*, *59*, 384–391. doi: 10.1177/00131649921969938

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* New York: Science Press.

Tucker, L. R., & MacCallum, R. C. (1997). Exploratory factor analysis. *Unpublished manuscript.*

u/kriesniem. (2018). *Different results for Pearson's R as compared to SPSS.* Retrieved from `https://www.reddit.com/r/rstats/comments/a7mzwo/different_results_for_pearsons_r_as_compared_to/`

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*, 219–246. doi: 10.1177/0095798418771807

Wrigley, C. (1957). The distinction between common and specific variance in factor theory. *British Journal of Statistical Psychology*, *10*, 81–98. doi: 10.1111/j.2044-8317.1957.tb00180.x

Ximénez, C. (2006). A monte carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling*, *13*(4), 587–614. doi: 10.1207/s15328007sem1304_5

Ximénez, C. (2009). Recovery of weak factor loadings in confirmatory factor analysis under conditions of model misspecification. *Behavior Research Methods*, *41*(4), 1038–1052. doi: 10.3758/BRM.41.4.1038

Yarkoni, T. (2019). *The generalizability crisis*. PsyArXiv. doi: 10.31234/osf.io/jqw35

Appendix

Table A1

Differences Between the R psych and SPSS Implementations

Procedure	Setting	R psych	SPSS	Note
PAF	Communality method	SMC, if this fails: unity	SMC, if this fails: MAC, if this fails, unity	How the diagonal of the original matrix is replaced to find initial eigenvalues. .
	Absolute eigenvalues	No	Yes	To avoid negative eigenvalues, SPSS takes the absolute of initial eigenvalues. This is not done in R psych, where negative eigenvalues might render the use of SMCs impossible.
	Criterion type	Difference in sum of communalities	Difference in maximum individual communalities	Value on which the convergence criterion is applied.
Promax	Varimax type	Singular value decomposition	Kaiser	SPSS follows the original varimax procedure from (Kaiser, 1958; likely with small changes in the varimax criterion), while R uses singular value decomposition.
	Normalization of target matrix	Unnormalized	Normalized	The target matrix is row-normalized in SPSS, but not in R psych. This is not the Kaiser normalization, which is done in both implementations.

Note: A detailed description of the implementations of R psych and SPSS can be found in the supplemental material. PAF = principal axis factoring; SMC = squared multiple correlation; MAC = maximum absolute correlation, unity = all 1's.

Table A2

Description of Population Models used for Simulation Analyses

a) Pattern Matrices

Pattern Matrices	m	Indicators per Factor	Size of Pattern Coefficients	Notes
Case 18|3|6	3	6	.60	Same baseline model as used in De Winter and Dodou (2012)
Case 6|3|6	3	2	.60	
Case 9|3|6	3	3	.60	Case 5 in De Winter and Dodou (2012)
Case 12|3|6	3	4	.60	
Case 15|3|6	3	5	.60	
Case 18|3|3	3	6	.30	
Case 18|3|9	3	6	.90	
Case 18|3|369b	3	6	.30, .60, .90	Different pattern coefficients between factors. Case 7 in De Winter and Dodou (2012)
Case 18|3|369w	3	6	.30, .60, .90	Different pattern coefficients within factors (each factor two each). Similar to cases 8/9 in De Winter and Dodou (2012)
Case 18|3|46|1c	3	6	.60	One cross-loading of .40. Similar to case 10 in De Winter and Dodou (2012)
Case 18|3|46|3c	3	6	.60	Three cross-loadings of .40 (One factor with 2 and one with 1 cross-loading). Similar to case 10 in De Winter and Dodou (2012)
Case 12|3m|6	3	2, 4, 6	.60	Similar to cases 11/ 12 in De Winter and Dodou (2012)
Case 18|3|6n	3	6	.60	Random variation in pattern coefficients added, drawn from a uniform distribution [-.2, .2]. Case 13 in De Winter and Dodou (2012)
Case 6|3|369wb	3	2	.30, .60, .90	Different pattern coefficients within one of the factors
Case 9|3|369wb	3	3	.30, .60, .90	Different pattern coefficients within and between factors
Case 12|3|369wb	3	4	.30, .60, .90	Different pattern coefficients within and between factors
Case 15|3|369wb	3	5	.30, .60, .90	Different pattern coefficients within and between factors
Case 12|6|6	6	2	.60	
Case 18|6|6	6	3	.60	
Case 24|6|6	6	4	.60	
Case 30|6|6	6	5	.60	
Case 36|6|6	6	6	.60	
Case 12|6|369wb	6	2	.30, .60, .90	Different pattern coefficients within and between factors

(continued)

Case 18|6|369wb	6	3	.30, .60, .90	Different pattern coefficients within and between factors
Case 24|6|369wb	6	4	.30, .60, .90	Different pattern coefficients within and between factors
Case 30|6|369wb	6	5	.30, .60, .90	Different pattern coefficients within and between factors
Case 36|6|369w	6	6	.30, .60, .90	Different pattern coefficients within factors

Note: A population model is always a combination of a population pattern matrix and a population factor intercorrelation matrix. m = Number of factors. All population models are available in the *EFAtools* package.

b) Factor Intercorrelations

Factor Intercorrelations	Size of Intercorrelations	Notes
Zero	.00	Same intercorrelations as used in De Winter and Dodou (2012)
Moderate	.30	
Mixed	.30, .50, .70	
Strong	.70	Same intercorrelations as used in De Winter and Dodou (2012)

Note: A population model is always a combination of a population pattern matrix and a population factor intercorrelation matrix. All population models are available in the *EFAtools* package.

Appendix E: Curriculum Vitae

Markus Steiner

Address	Hasenweid 14 4600 Olten
Telephone	+41 79 266 06 88
E-Mail	markus.d.steiner@gmail.com

Education

02.2018 – present	**PhD in Cognitive and Decision Science** Center for Cognitive and Decision Sciences University of Basel (Advisors: Prof. Dr. Rui Mata and Prof. Dr. Jörg Rieskamp) PhD thesis title: «Establishing construct validity: The cases of risk preference and exploratory factor analysis»
08.2016 – 01.2018	**MSc in Psychology** Social-, Economic-, and Decision Psychology master's program University of Basel (Advisor: Dr. Nathaniel Phillips) Master's thesis title: «Getting what you came for: Decisions under uncertainty with a goal»
08.2013 – 07.2016	**BSc in Psychology** University of Basel (Advisor: Prof. Dr. David Kellen) Bachelor's thesis title: «Two Theories of Risky Decision Making and Their Relation to Numerical Competences»

Academic Experience

01.2018 – present	**Assistant/PhD Student** University of Basel • Plan, program, conduct, and analyze studies • Write scientific articles • Present research findings at international conferences • Plan and teach workshops and seminars
09.2017 – present	**Instructor** The R Bootcamp, Basel • Plan, prepare, and teach workshops on data science, statistics, machine learning, and reporting with R • Author blog entries for the webpage
07.2017 – 09.2017 06.2015 – 09.2015 01.2015 – 02.2015	**Intern** University of Basel and University Psychiatric Clinics, Basel Multiple scientific internship, including writing the *ShinyPsych* R package
09.2014 – 12.2017	**Research Assistant** University of Basel • Program and conduct experiments • Statistics tutoring • Operate MRI in a neuroscience study

Publications

Submitted articles

- Grieder, S. & **Steiner, M. D.** (2020). *Algorithmic jingle jungle: A comparison of implementations of principal axis factoring and promax rotation in R and SPSS.* Manuscript submitted for publication. https://doi.org/10.31234/osf.io/7hwrm

Accepted/published articles (peer-reviewed)

- **Steiner, M.D.** & Frey, R. (in press). Representative design in psychological assessment: A case study using the balloon analogue risk task (BART). *Journal of Experimental Psychology: General*. Retrieved from https://psyarxiv.com/dg4ks/
- **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*. Retrieved from https://psyarxiv.com/sa834/
- **Steiner, M. D.** & Grieder, S. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53), 2521. https://doi.org/10.21105/joss.02521
- Kellen, D., **Steiner, M.D.**, Davis-Stober, C.P., & Pappas, N.R. (2020). Modeling choice paradoxes under risk: From prospect theories to sampling-based accounts. *Cognitive Psychology*, 118, 101258. https://doi.org/10.1016/j.cogpsych.2019.101258
- Mueller, F., Lenz, C., **Steiner, M.D.**, Dolder, P.C., Walter, M., Lang, U.E., Liechti, M.E., Borgwardt, S. (2016). Neuroimaging in moderate MDMA use: A systematic review. *Neuroscience & Biobehavioral Reviews*, *62*, 21-34. https://doi.org/10.1016/j.neubiorev.2015.12.010

Conferences

2020 Seitz, F.I., **Steiner, M.D.**, & Frey, R., Through the window of my mind: Mapping information integration and the cognitive representations underlying self-reported risk preference, SJDM 41st Annual Conference, December 9 – 12, *poster*

2020 **Steiner, M.D.**, & Frey, R., Representative design in psychological assessment: A case study using the Balloon Analogue Risk Task (BART), SJDM 41st Annual Conference, December 9 – 12, *talk*

2020 **Steiner, M.D.**, & Frey, R., Representative design in psychological assessment: A case study using the Balloon Analogue Risk Task (BART), 26th International (Virtual) Meeting of the Brunswik Society, December 3 – 4, *talk*

2020 **Steiner, M.D.**, Seitz, F.I., & Frey, R., Through the window of my mind: Mapping information integration and the cognitive representations underlying self-reported risk preference, Virtual Process Tracing Conference, September 17, *talk*

2019 **Steiner, M.D.**, Seitz, F.I., & Frey, R., Mapping the cognitive processes underlying self-reported risk-taking propensity, SJDM 40th Annual Conference, November 15 – 18, *poster*

2019 **Steiner, M.D.**, Seitz, F.I., & Frey, R., Mapping the cognitive processes underlying self-reported risk-taking propensity, SPUDM, August 18 – 22, *poster*

2019 **Steiner, M.D.**, Seitz, F.I., & Frey, R., Mapping the cognitive processes underlying self-reported risk-taking propensity, TeaP Conference, April 15 –17, *talk*

2018 **Steiner, M.D.**, & Frey, R., Beyond risk preference? Modeling risk taking with state-specific mechanisms, SJDM 39th Annual Conference, November 16 – 19, *poster*

2018 **Steiner, M.D.**, & Frey, R., Beyond "risk preference"? Towards a general model of risk-taking behavior, Summer Institute on Bounded Rationality, June 19 – 27, *poster*

2018 **Steiner, M.D.**, & Frey, R., Beyond "risk preference"? Towards a general model of risk-taking behavior, 11th JDMx meeting for Early Career Researchers, June 6 – 8, *talk*

Academic Skillset

R:	Proficient (author of the *EFAtools* and *ShinyPsych* R packages that implement exploratory factor analyses and facilitate programming studies in R shiny, respectively; applying and teaching R for data preparation, analysis, visualization, and reporting)
SQL:	Intermediate (programming online studies, data wrangling)
HTML, JS, CSS, PHP:	Intermediate (programming online studies)
Python	Intermediate (programming studies, workshops on data wrangling)
LaTeX:	Intermediate (writing scientific articles)
MS Excel/ Google Sheets:	Intermediate (advanced formulas, pivot tables)