frontiers in Psychology

MINI REVIEW
published: 21 November 2018
doi: 10.3389/fpsyg.2018.02307

Can We Infer Inter-Individual Differences in Risk-Taking From Behavioral Tasks?

Stefano Palminteri[1,2,3]* *and Coralie Chevallier*[1,2,3]

[1] *Laboratoire de Neurosciences Cognitives, Institut National de la Santé et de la Recherche Médicale, Paris, France,* [2] *Département d'Etudes Cognitives, Ecole Normale Supérieure, Paris, France,* [3] *Institut d'Etudes de la Cognition, Université de Paris Sciences et Lettres, Paris, France*

Investigating the bases of inter-individual differences in risk-taking is necessary to refine our cognitive and neural models of decision-making and to ultimately counter risky behaviors in real-life policy settings. However, recent evidence suggests that behavioral tasks fare poorly compared to standard questionnaires to measure individual differences in risk-taking. Crucially, using model-based measures of risk taking does not seem to improve reliability. Here, we put forward two possible – not mutually exclusive – explanations for these results and suggest future avenues of research to improve the assessment of inter-individual differences in risk-taking by combining repeated online testing and mechanistic computational models.

Keywords: risk-taking, inter-individual variability, behavioral phenotype, behavioral economics, correlational psychology

OPEN ACCESS

Edited by:
Michael Banissy,
Goldsmiths, University of London, United Kingdom

Reviewed by:
Ariel Telpaz,
General Motors, United States
Joshua Weller,
Tilburg University, Netherlands

***Correspondence:**
Stefano Palminteri
stefano.palminteri@ens.fr

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *28 May 2018*
Accepted: *05 November 2018*
Published: *21 November 2018*

Citation:
Palminteri S and Chevallier C (2018) Can We Infer Inter-Individual Differences in Risk-Taking From Behavioral Tasks? Front. Psychol. 9:2307. doi: 10.3389/fpsyg.2018.02307

INTRODUCTION

In a recent series of studies Frey et al. (2017) investigated the relationship between different measures of risk-sensitivity in a laboratory-based experiment involving over a thousand participants (N∼1500) (Frey et al., 2017; Pedroni et al., 2017). By comparing standard behavioral tasks, personality questionnaires, and reports of actual frequency of risky behaviors, the authors were able to demonstrate that behavioral tasks are consistently less reliable than questionnaires. First, performance in risk-taking tasks were less correlated to actual frequency of risky behaviors than personality scores, which suggests that external validity is low. Second, behavioral measures were less correlated among themselves than personality scores and frequency measures, which suggests that they tap constructs that are less consistent (low between-task reliability). These findings are not isolated: other studies from other laboratories, involving smaller number of subjects and behavioral tasks, reached very similar conclusions (Corsetto and Filippin, 2013; Attanasi et al., 2018). Beyond raw behavioral measures, a computational modeling approach using cumulative prospect theory (CPT) parameters (decreasing marginal utility, loss aversion and subjective weighting of probabilities) failed to improve between-task reliability. Finally, test-retest reliability was lower for behavioral tasks than for personality scores. Strikingly, preliminary evidence suggests that these issues generalize to many behavioral tasks beside decision-making under risk, such as reinforcement learning (Enkavi et al., 2018). These findings are not isolated: other studies from other laboratories, involving smaller number of subjects and behavioral tasks, reached very similar conclusions (Corsetto and Filippin, 2013; Attanasi et al., 2018).

Low external validity and-reliability is extremely worrying for the development of behavioral economics applications and (by extension) for the neuroeconomics research framework, where risk preferences are commonly assessed and elicited using behavioral tasks. In addition, the unreliability

of behavioral measures is also problematic for the computational psychiatry research framework that has recently placed great emphasis on the use of cognitive and behavioral phenotyping tools. The idea behind these frameworks is that behavioral measures can be used to phenotype patients at the individual level and ultimately work as tools to perfect diagnosis, personalize care, and assess the efficacy of new treatments or drugs in clinical trials (Huys et al., 2016). In this context, it is therefore vital that behavioral tasks generate results that are stable and predictive of real life outcomes.

In addition to questioning approaches based on behavioral phenotyping tools, these findings also raise a profound epistemological challenge. Given that real life frequency of risky behaviors is the reflection of past choices, why then, do personality measures – that are based on *questionnaires* – explain real life behaviors better than behavioral measures – that are based on *choices*? And why would the same subjects produce different choices when presented with the very same task twice?

TWO POSSIBLE EXPLANATIONS

We put forward two possible answers for these puzzling results and fundamental questions (low external validity and consistency of behavioral measures): The first possibility is that these findings reflect a **problem with the instrument;** The other possibility is that these findings reflect a **problem with the construct**.

The "**problem with the instrument**" argument has been explicitly put forward by the authors of the studies (Frey et al., 2017; Pedroni et al., 2017). According to this hypothesis, the low external validity and reliability of the behavioral tasks derive from intrinsic limitations of the tasks. For instance, it has been argued that low between-task consistency between behavioral measures derives from the fact that each task involves both central (risk sensitivity) and peripheral processes (responses, stimuli), whose variability may affect the results. Low test-retest reliability should also be expected given that behavioral and cognitive tasks are traditionally designed to reduce between-subjects variance and to maximize between-conditions variance, such that the very features that make a behavioral task "successful" (high reproducibility of the "average" results) make it unsuited to assess inter-individual differences (Hedge et al., 2017). As nicely summarized by Hedge et al. (2017):

> *"Experimental effects become well established – and thus those tasks become popular – when between-subject variability is low. However, low between-subject variability causes low reliability for individual differences, destroying replicable correlations with other factors and potentially undermining published conclusions drawn from correlational relationships."*

Propensity measures on the other hand, are designed to maximize inter-individual differences. In addition, a good test-retest reliability is a *condicio sine qua non* for the publication of personality questionnaires, hence their good temporal consistency. Finally, in the context of the specific set of studies at hand, it is also worth noting that the frequency measures were assessed using self-report questionnaires, which involve the same response modality as the personality measures. Furthermore, risk propensity and risk frequency assessments shared similar content and it should come as no surprise that subjects provide similar responses to similar questions, e.g., in order to present a coherent image of themselves (a good "narrative"). In statistical terms, this would result in an artificially increased correlation between frequency and personality measures. Taken together, these features may inflate the consistency and validity of the personality measures. Finally, self-reported questionnaires are well-known for eliciting edulcorated representations that are influenced by a range of social norms (Edwards, 1953). To overcome the issues raised by self-reported frequency of risk behaviors, personality and behavioral measures should be tested against objective assessments of risky behaviors (e.g., expired CO_2 for smoking, medical records, etc.).

The argument that there is a "problem with the instrument" also applies to the mathematical model used to quantify risk propensity parameters. The authors indeed focused on CPT, which is a widely used descriptive model originally designed to explain one-shot decisions. But three features of CPT may undermine the internal consistency of model-based measures of risk sensitivity (Tversky and Kahneman, 1992). First, different tasks engage different peripheral processes but the same CPT model is applied to various behavioral tasks with no task-specific adjustment of the functional form. Second, and more importantly, CPT parameters are assumed to be static and not affected by the individual's history of choice, by relevant contextual factors or by feedback. In that respect, CPT is a purely descriptive model rather than a mechanistic model. Third, CPT parameters are often correlated and it is often hard to disentangle their respective contribution to risky behavior using standard fitting procedures. This is in part because different parameter values can produce the same behavioral phenotype (e.g., loss aversion) (Nilsson et al., 2011), which may undermine the power of the model to unambiguously predict particular behavioral profiles.

The "**problem with the construct**" argument implies that behavioral tasks provide a genuine estimate of the subject's momentary risk attitude at the time of testing, but that risk attitude itself changes over time. This is plausible if we assume that momentary risk attitude is influenced by multiple factors. To illustrate this idea, we now consider a simplified case involving two possible phenotypes, a risk-seeking phenotype (red) and a risk-averse phenotype (blue), and we propose a multi-layer model in which momentary risk attitude corresponds to the weighted sum of different sources of influence that change with different time constants (**Figure 1**). In this toy example, the first layer corresponds to the subject's "trait," which is determined by her genotype and which remains stable over her lifespan. The last layer corresponds to random (or unpredictable) factors, such as unexpected external stimuli and contextual factors. In between these two extremes, we hypothesize that additional sources of influence are at play, such as very slow age-related changes and very fast circadian rhythms. According to this model, a subject tested twice with the same behavioral task at different time points will not necessarily display the same phenotype. Within

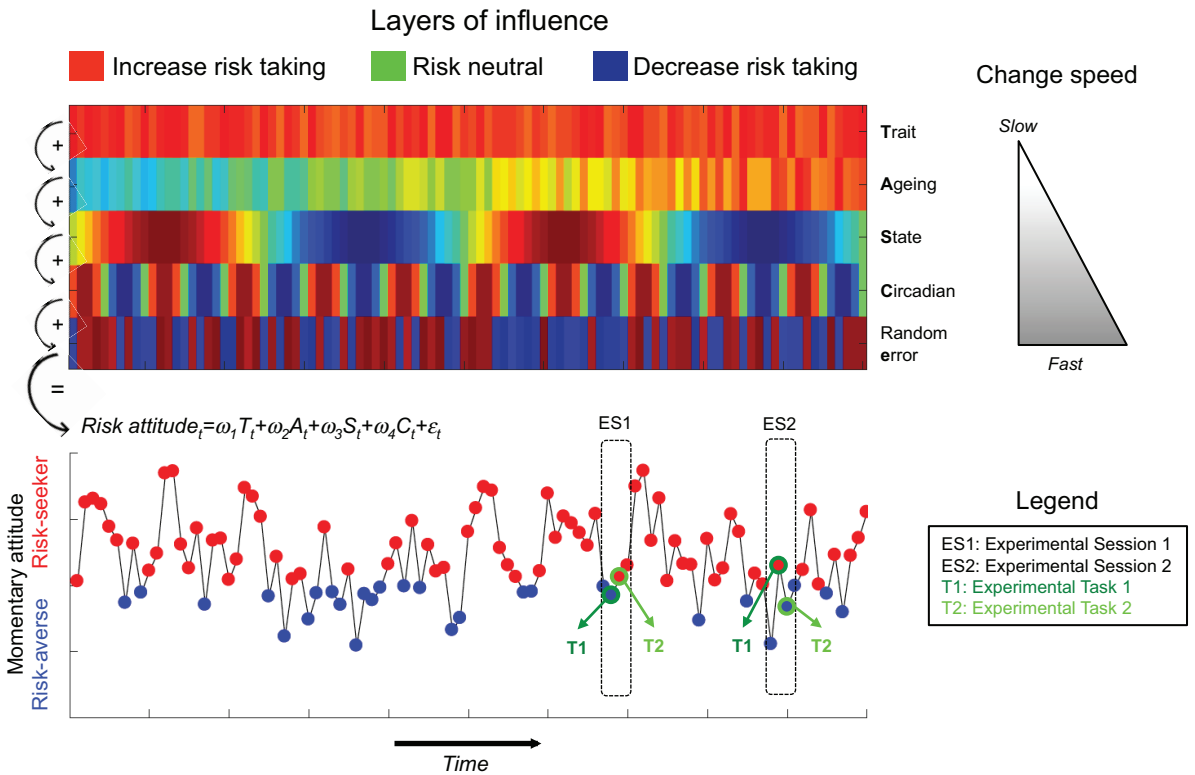

FIGURE 1 | The figure schematizes how low consistency of behavioral measures of risk may arise from the multi-layer model. At the top, we represent the different factors that influence the probability to express a given behavioral phenotype at a given time point in addition to random error. We consider a simplified case in which only two phenotypes are possible: red (risk seeking) and blue (risk aversion). The different layers change at different time constants (as exemplified by the gray triangle on the right). At a given time point (t) the momentary risk attitude is the weighted sum of the different layers of influence plus random error. A given subject is tested in two experimental sessions (ES1, and ES2) with two behavioral tasks supposed to measure the same behavioral phenotype (T1 and T2). The multi-layer model may explain why behavioral measures are not consistent between-tasks and between-sessions.

this framework, the fact that propensity measures produce more stable results can be explained by the fact that filling out questionnaires relies on cognitive processes that do not involve risk attitude *per se*, such as robust averaging of previous experiences stored in episodic memory or introspection.

Crucially, there is evidence demonstrating that these various layers are indeed relevant to understanding decision-making under risk: genetic factors influence risk-related behaviors (Linner et al., 2018), behavioral measures of risk sensitivity evolve across the life-span (Weller et al., 2011), and are affected by hormonal and circadian factors (Lazzaro et al., 2016; Glimcher and Tymula, 2017), mood states (Stanton et al., 2014), as well as momentary arousal (FeldmanHall et al., 2016). Importantly, the same factors are involved in other decision-making processes such as cooperation in economics dilemmas, a field where behavioral tasks also predict real life behaviors poorly (Gurven and Winking, 2008; Winking and Mizer, 2013). By contrast, propensity measures, as implemented by questionnaires, are designed to assess participants' prototypical behavior averaged across long period of times, thus canceling out momentary trends. In other words, questionnaires are designed to assess stable "*traits*." In many cases, participants are explicitly instructed to extract their prototypical behavior with formulations such as "*describe yourself as you generally are*" and to ignore the variability induced by circadian or age-related changes.

CONCLUSION AND PERSPECTIVES

Recent evidence based on large-scale behavioral testing shows that behavioral measures in cognitive tasks are outperformed by propensity measures from personality questionnaires, in terms of external validity (i.e., correlation with frequency measures) and reliability (between-tasks consistency and test-retest reliability). We delineate two possible – not mutually exclusive – interpretations of these results. The pessimistic "problem with the instrument" argument states that behavioral tasks are not suited to investigate inter-individual differences. The optimistic "problem with the construct" argument states that variability in behavioral tasks reflects true changes in momentary risk attitude. According to this view, behavioral tasks reflect true momentary risk attitude and will the quantification of the relative weights of the different layers.

At the moment, personality questionnaires appear to be the best psychological tools to predict the frequency of real-life risky behavior. Should we then, abandon the quest for behavioral

measures of individual variability? Probably not. Questionnaires are hugely informative when it comes to providing an accurate description of the variability with which personality traits manifest but they cannot be used to trace back the cognitive and neural mechanisms that together produce such variability. The paucity of robust behavioral tools to characterize inter-individual differences therefore constitutes an important obstacle in building proper models of cognitive variability.

Developing behavioral biomarkers, however, requires a proper re-think in the way cognitive scientists design tasks so that they maximize between-subjects variance. One promising possibility is to shift from fixed and passive designs to **active and adaptive** ones. Adjusting task parameters online could indeed correct for momentary changes in baseline performance that may affect the assessment of risk preferences. These results also highlight the importance of **repeated testing**, which has now become considerably easier with the development of smart-phone based behavioral experiments. Repeated testing should also allow us to test the multi-layer hypothesis, to attribute precise coefficients to the different layers, and by averaging performance over experiments, to infer participants' trait-level phenotype. The issue related to the ambiguous relationship between CTP parameters and behavioral profiles and their correlation may be solved by implementing principal component analyses instead of working with the raw parameters and by implementing hierarchical model fitting (Nilsson et al., 2011). This approach would of course require external validation to assess which component reflects risk sensitivity but we believe it is a valuable alternative to current methods. Ultimately, developing and refining **mechanistic and dynamic models** of risk preferences that integrate learning processes and contextual factors, might also allow for a better quantification of risk preferences at the individual level. A promising way to design these models could be the development of choice prediction competitions, a method that already commonly used in the machine learning literature (Erev et al., 2017). Even more ambitiously, these prediction competitions would include data collection at multiple time points as well as external validation by real life outcomes.

AUTHOR CONTRIBUTIONS

SP designed the review. SP and CC wrote the review.

FUNDING

SP was supported by an ATIP-Avenir grant (R16069JS) Collaborative Research in Computational Neuroscience ANR-NSF grant (ANR-16-NEUC-0004), the Programme Emergence(s) de la Ville de Paris, and the Fondation Fyssen. The Institut d'Etudes de la Cognition was supported financially by the LabEx IEC (ANR-10-LABX-0087 IEC) and the IDEX PSL* (ANR-10-IDEX-0001-02 PSL*).

ACKNOWLEDGMENTS

We thank Nathaniel Daw for useful comments.

REFERENCES

Attanasi, A., Georgantzís, N., Rotondi, V., and Vigani, D. (2018). Lottery- and survey-based risk attitudes linked through a multichoice elicitation task. *Theory Decis.* 84, 341–372. doi: 10.1007/s11238-017-9613-0

Corsetto, P., and Filippin, A. (2013). A theoretical and experimental appraisal of five risk elicitation methods. *SOEPpapers on Multidisciplinary Panel Data Research 547*, Berlin. doi: 10.2139/ssrn.2253819

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *J. Appl. Psychol.* 37, 90–93. doi: 10.1037/h0058073

Enkavi, A., Eisenberg, L., Bissett, P., Mazza, G. L., Mackinnon, D. P., Marsch, L. A., et al. (2018). A large-scale analysis of test-retest reliabilities of self-regulation measures. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/x5pm4

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

FeldmanHall, O., Glimcher, P., Baker, A. L., and Phelps, E. A. (2016). Emotion and decision-making under uncertainty: physiological arousal predicts increased gambling during ambiguity but not risk. *J. Exp. Psychol. Gen.* 145, 1255–1262. doi: 10.1037/xge0000205

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., and Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Sci. Adv.* 3:e1701381. doi: 10.1126/sciadv.1701381

Glimcher, P. W., and Tymula, A. (2017). Let the sunshine in: the effects of luminance on economic preferences, choice consistency and dominance violations. *PLoS One* 12:e0181112. doi: 10.1371/journal.pone.0181112

Gurven, M., and Winking, J. (2008). Collective action in action: prosocial behavior in and out of the laboratory. *Am. Anthropol.* 110, 179–190. doi: 10.1111/j.1548-1433.2008.00024.x

Hedge, C., Powell, G., and Sumner, P. (2017). The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50, 1166–1186. doi: 10.3758/s13428-017-0935-1

Huys, Q. J., Maia, T. V., and Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* 19, 404–413. doi: 10.1038/nn.4238

Lazzaro, S. C., Rutledge, R. B., Burghart, D. R., and Glimcher, P. W. (2016). The impact of menstrual cycle phase on economic choice and rationality. *PLoS One* 11:e0144080. doi: 10.1371/journal.pone.0144080

Linner, R. K., Biroli, P., Kong, E., Meddens, S. F. W., Wedow, R., Fontana, M. A., et al. (2018). Genome-wide study identifies 611 loci associated with risk tolerance and risky behaviors. *bioRxiv* [Preprint]. doi: 10.1101/261081

Nilsson, H., Rieskampa, J., and Wagenmakers, E. (2011). Hierarchical bayesian parameter estimation for cumulative prospect theory. *J. Math. Psychol.* 55, 84–93. doi: 10.1016/j.jmp.2010.08.006

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

Stanton, S. J., Reeck, C., Huettel, S. A., and LaBar, K. S. (2014). Effects of induced moods on economic choices. *Judgem. Decis. Mak.* 9, 167–175.

Tversky, A., and Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* 5, 297–323. doi: 10.1007/BF00122574

Weller, J. A., Levin, I. P., and Denburg, N. L. (2011). Trajectory of risky decision making for potential gains and losses from ages 5 to 85. *J. Behav. Decis. Mak.* 24, 331–344. doi: 10.1002/bdm.690

Winking, J., and Mizer, N. (2013). Natural-field dictator game shows no altruistic giving. *Evol. Hum. Bahav*. 34, 288–293. doi: 10.1016/j.evolhumbehav.2013.04.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.