Gaertner, Christine; Steinorth, Petra

On the correlation of self-reported and behavioral risk attitude measures: The case of the General Risk Question and the Investment Game following Gneezy and Potters (1997)

Risk Management and Insurance Review

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Gaertner, Christine; Steinorth, Petra (2023) : On the correlation of self-reported and behavioral risk attitude measures: The case of the General Risk Question and the Investment Game following Gneezy and Potters (1997), Risk Management and Insurance Review, ISSN 1540-6296, Wiley, Hoboken, NJ, Vol. 26, Iss. 3, pp. 367-392, https://doi.org/10.1111/rmir.12250

This Version is available at:
https://hdl.handle.net/10419/288217

DOI: 10.1111/rmir.12250

FEATURE ARTICLE

Risk Management and Insurance Review WILEY

On the correlation of self-reported and behavioral risk attitude measures: The case of the General Risk Question and the Investment Game following Gneezy and Potters (1997)

Christine Gaertner | **Petra Steinorth**

Universität Hamburg, Hamburg, Germany

Correspondence
Petra Steinorth, University of Hamburg, Moorweidenstrsse 18, Hamburg 20148, Germany.
Email: petra.steinorth@uni-hamburg.de

Funding information
Deutsche Forschungsgemeinschaft (DFG), Grant/Award Number: 433254283

Abstract

Risk attitudes play a pivotal role to understand economic decision-making, and several measures are used to elicit them in the lab and survey them in the field. We provide a literature review on the most commonly used risk elicitation methods by Holt and Laury (HL) and the Investment Game (IG) by Gneezy and Potters and the General Risk Question (GRQ) utilized in the German Socioeconomic Panel. Based on the metadata from three experiments, we show that the GRQ has a robust and economically relevant association with the IG.

1 | INTRODUCTION

Many decisions in everyday life include risk (Mata et al., 2018). For example, the decision how to invest money or changing jobs. The basis for understanding and predicting behavior is a person's underlying risk attitude (Charness et al., 2013, 2020; Dohmen et al., 2011). This can be important in different contexts: For example, financial institutions such as banks or insurance companies should adapt their products and services to the risk attitudes of their customers or target group (Holt & Laury, 2014; Pedroni et al., 2017). Also, researchers may make use of risk attitudes to control for different risk preferences when investigating topics that involve some kind of risk (Dohmen et al., 2011).

To make risk attitudes tangible, researchers have developed numerous elicitation methods over the past three decades. Thereby, the methods can be divided into two main categories: (1) Self-report measures, where individuals are directly asked to assess their own risk attitude. These are more easily administered in large surveys. (2) Methods are behavioral measures, where risk attitudes are derived from observed behavior in (usually) incentivized tasks (Crosetto & Filippin, 2013a). While self-report measures are typically presented in the form of questionnaires, behavioral measures are mostly presented in the form of lotteries or in a game-style character. They differ with respect to the number of decisions that have to be made (Crosetto & Filippin, 2013a; Dasgupta et al., 2019). Behavioral measures are often incentivized with a final payoff and are more expensive than self-report measures (Lönnqvist et al., 2015). In addition, while complex tasks might generate more precise results, they may also lead to comprehension problems (Charness et al., 2013).

Given the variety of measures, one pressing question is the association of the different risk attitude measures, which should be statistically significant and economically relevant as they all measure risk attitude (Attanasi et al., 2018; Pedroni et al., 2017). However, weak relations between risk attitude measures are a recurrent finding in literature (see e.g., Charness & Viceisza, 2011; Frey et al., 2017; Galizzi & Miniaci, 2016; Lönnqvist et al., 2015). The study of Pedroni et al. (2017) illustrates this strikingly for behavioral risk measures, as the share of subjects classified as risk-averse ranges between 21% and 100% in different tasks. Such substantial differences put the reliability of the measures into question, which is important especially when risk attitudes are elicited for real-world business purposes (e.g., when allocating capital in risky and risk-free asset options according to the customer's elicited risk attitude; Dasgupta et al., 2019; Pedroni et al., 2017).

This paper addresses the question how the commonly used self-reported measures of the General Risk Question (GRQ)[1] relate to the commonly used behavioral measure, the Investment Game (IG) by Gneezy and Potters (1997). The IG is a lottery task involving only one decision. Subjects receive an endowment and have to determine which share of it they are willing to invest in a risky asset. If the risky asset develops positively, a multiple of the investment is paid out; otherwise, the invested share is lost (Gneezy & Potters, 1997).

While several studies have investigated the correlation of the SOEP questions with the HL task, evidence on the correlation with the IG is scarce. To the best of the authors knowledge, the latter relation has only been investigated in the study of Crosetto and Filippin (2013a) and Menkhoff and Sakha (2017). Both find no significant relation between the two measures. The former study used a relatively small sample of only 86 subjects, which may have impacted the statistical power (Wolf & Best, 2010). Also, both studies used relatively low endowments (€4 and 100 Thai Bhat [approximately €2 at that time], respectively), which might have been too low to motivate subjects to disclose true risk preferences. Our paper contributes to the discussion on the association of the GRQ with the IG by analyzing three larger samples (115, 325, and 420 subjects) with higher IG endowments (€7–€10.13). In contrast to previous work, we find a statistically significant and economically meaningful association of these two measures in the metadata from three economic experiments. One novelty of this paper is that two different versions of the IG are incorporated into the analysis. In the original version of the IG, the expected return of investing is positive, making an investment attractive even for risk-averse individuals. Additionally, a version of the IG is investigated, where the expected return

[1]As administered in the German Socioeconomic Panel (SOEP) since 2004.

of investing is negative, which should lead to a different investment behavior. The research question pursued with the empirical analysis is how the GRQ is associated with both IG versions.

The findings of the empirical analysis can be summarized as follows: First, in contrast to previous studies, a reasonable significant and positive association between the GRQ scores and the risky shares in the IG is found. This holds for both versions of the IG.

After this introduction, we first provide a thorough literature review on commonly utilized self-reported and the most commonly utilized behavioral risk measures following Holt and Laury (2002) and the IG following Gneezy and Potters (1997). Section 3 summarizes the three experiments and shows results of the pooled data as well as each of the experiment. We conclude with a brief discussion.

2 | LITERATURE REVIEW

2.1 | Strengths and weaknesses of the GRQ, HL, and the IG

This section reviews the strengths and weaknesses of the SOEP risk questions, the HL task and the IG in terms of implementation and time effort, involved costs, complexity of the task, spectrum and preciseness of elicitable risk attitudes, and temporal stability. Table 1 summarizes the discussion of this section.

All three measures perform well in terms of implementation and time effort. It does not take much time to prepare the SOEP questionnaire, the table from the HL task, or the task instructions of the IG, as all tasks can be executed relatively quickly with pen and paper or with a computer (Crosetto & Filippin, 2013a; Maier & Rüger, 2010; Vieider et al., 2013).

Selecting a score in the SOEP questionnaire does not require much time. Although respondents have to put a little more thought into their decisions in the HL task and in the IG, as probabilities and payoffs are involved, the required time is still limited (Crosetto & Filippin, 2013a). One aspect that causes the effort in the HL task and the IG to be somewhat higher than in the SOEP questions is that compensation is usually required. The elicitation of risk attitudes with the HL task and the IG is associated with a payout for every subject

TABLE 1 Strengths and weaknesses of the SOEP questions, HL task, and IG.

	SOEP risk questions	HL task	IG
Implementation effort	Low	Low to medium	Low to medium
Time effort	Low	Low to medium	Low to medium
Costs	Low	Medium to high	Medium to high
Comprehensibility	Good	Poor in some groups	Good
Spectrum of risk attitudes	Unlimited	Unlimited	Limited
Risk-aversion coefficient	Calculation not possible	Calculation possible	Calculation possible
Precision	-	Intervals	Almost continuous
Temporal stability	Strong	Poor	Poor

Abbreviations: HL, Holt and Laury; IG, Investment Game; SOEP, Socioeconomic Panel.

(Dohmen et al., 2011; Lönnqvist et al., 2015). The costs vary, depending on the chosen stakes in the HL task and the IG.[2]

With respect to perceived complexity of the tasks, there are considerable differences. While the SOEP risk questions do not require much explanation, the explanatory effort is higher in the HL task and the IG due to the involved probabilities and different possible outcomes (Dasgupta et al., 2019; Dave et al., 2010). Interestingly, researchers found that the SOEP questions and the IG are relatively easy for subjects to understand (Charness et al., 2013, 2020; Charness & Viceisza, 2011; Crosetto & Filippin, 2013a; Dasgupta et al., 2016; Holt & Laury, 2014; Lönnqvist et al., 2015; Verschoor et al., 2016). In Crosetto and Filippin (2013a), subjects rated the IG as comparatively simple to other tasks. Also, data from rural populations in Senegal and Uganda showed no signs that the IG was not understood (Charness & Viceisza, 2011; Verschoor et al., 2016). Regarding the comprehensibility of the HL task, researchers are divided: While the HL task exhibited no comprehension problems when it was piloted in the rural Ugandan population (Verschoor et al., 2016), 75% of answers were inconsistent in a sample from rural Senegal (Charness & Viceisza, 2011). Dave et al. (2010) also experienced increased comprehension questions and inconsistent responses in the HL task, especially among people with a lack of mathematical understanding. Some researchers criticize that the original HL task allows inconsistent answers as it is possible to switch back and forth between the two lotteries (Andersen et al., 2006; Charness et al., 2013; Crosetto & Filippin, 2013a). To prevent this, Andersen et al. (2006) propose to either offer an option where subjects can indicate that they are indifferent between both lotteries or to ask subjects for the row where they would switch from lottery A to lottery B, instead of letting them decide in every row. Charness et al. (2013) warn that constraining consistent decisions might distort the results.

Another aspect worth considering is the spectrum and precision of elicitable risk attitudes. A strength of the HL task is that they cover the entire spectrum of risk attitudes, from risk averse (choosing lottery A for more than the first four rows in the HL task) to risk seeking (choosing lottery A for less than the first four rows) (Crosetto & Filippin, 2013a). Ranking subjects based on their SOEP scores should be treated with caution, as subjects might interpret the scores differently. For example, a score of 3 may indicate a higher level of risk aversion for one person than for another, as only the scores 0 and 10 are clearly defined. In the IG, the range of elicitable risk attitudes is limited, as IGs with positive (negative) expected return cannot disentangle risk-neutral from risk-seeking (risk-averse) individuals (both should invest the whole endowment [invest nothing]) (Charness et al., 2013; Charness & Viceisza, 2011; Crosetto & Filippin, 2013a; Dasgupta et al., 2016). With respect to the precision of elicited risk-aversion coefficients, the IG performs best: Due to the freedom of subjects to choose whatever share they wish to invest, the risk-aversion coefficients can be estimated almost continuously (Charness & Viceisza, 2011; Crosetto & Filippin, 2013a). In the HL task, only intervals of risk-aversion coefficients can be estimated (Andersen et al., 2006; Crosetto & Filippin, 2013a).[3] Andersen et al. (2006) propose to reduce this problem by shrinking the interval iteratively: After selecting

[2]However, the stakes should be selected with caution, as it was found that the ranges and values used in the HL task (and likely also in the IG) impact how subjects behave (Galizzi & Miniaci, 2016; Holt & Laury, 2002; Maier & Rüger, 2010). For example, Holt and Laury (2002) and Maier and Rüger (2010) find that higher stakes in the HL task increase risk aversion.

[3]For an example on how the risk-aversion coefficients in the HL task and the IG can be calculated, see Appendix A. The examples show that the range for the coefficient of risk aversion is smaller in the IG than in the HL task.

a switching point in the original HL table, the table can be refined and subjects can choose a new switching point within the refined interval (Andersen et al., 2006). In contrast, the scores from the SOEP questionnaire do not allow to calculate coefficients of risk aversion (Charness et al., 2020). Since risk-aversion coefficients are usually calculated by equating the expected utility of two options (containing payoffs with certain probabilities) to which the subject is indifferent, the SOEP questions lack information to calculate such a coefficient (Anderson & Mellor, 2009; Verschoor et al., 2016).

Another aspect is the temporal stability of the three measures: The subjects' decisions differ substantially when they repeat the tasks after a certain time. Researchers have found that risk attitudes elicited by the SOEP risk questions are stable over different time horizons (intervals up to 10 years) (Drichoutis & Vassilopoulos, 2021; Galizzi & Miniaci, 2016; Lönnqvist et al., 2015; Mata et al., 2018). A rather poor temporal stability is found for the HL task and the IG. Lönnqvist et al. (2015) conclude that their data of the HL task exhibits no test–retest stability over a time interval of 1 year, and Galizzi and Miniaci (2016) note that less than 33% of participants choose the same switching point 1 year later. With respect to the IG, Holden and Tilahun (2022) find a correlation of 0.135 for investment decisions made 1 year apart. They attribute the low test–retest stability to a measurement error in the IG as the changes cannot be explained by changes in the risk tolerance of real investments (Holden & Tilahun, 2022).

2.2 | Behavioral validity of the GRQ, HL, and the IG

In this section, we summarize findings from the literature on behavioral validity of the GRQ, HL, and the IG. In HL and in the IG, elicited risk attitudes are based on observations of real behavior (Andersen et al., 2006; Crosetto & Filippin, 2013a; Galizzi & Miniaci, 2016; Verschoor et al., 2016). Subjects are expected to act according to true risk preferences (Andersen et al., 2006). Yet, the evidence is rather mixed on whether behavioral measures outperform self-reported measures in terms of behavioral validity. There are two main approaches to evaluate the behavioral validity of the measures: The first approach is to let subjects complete an incentivized experimental task involving risk, to observe whether subjects behave according to their elicited risk attitude. The second approach is to compare the elicited risk attitudes with stated real-world risky behaviors. Table 2 summarizes the findings on the behavioral validity that are presented below.

Different kinds of experimental risky tasks have been used, to evaluate the behavioral validity with the first approach: Dohmen et al. (2011) and Vieider et al. (2013) conducted a real-stakes lottery experiment in a German sample and in samples from over 30 countries, respectively, and find that the behavior in the experiment is correlated with the choices in the SOEP risk questions. Further, Lönnqvist et al. (2015) identify that the SOEP measure is associated with choices in a trust game involving an investment decision, while no such relation was found with the HL task. Charness et al. (2020) find that the SOEP questions, the HL task, and the IG are associated with decisions in an experimental portfolio and mortgage task, while the insurance task they administer is only associated with the HL task. In sum, each measure is positively related to behavior in at least one of the experimental tasks.

With respect to self-reported measures, Dohmen et al. (2011) find that the context-specific risk questions are the best predictors for reported risky decisions in the corresponding context (they investigated reported behavior with respect to holding stocks, being self-employed, participating in sports, and smoking). For example, the financial-context risk question is the

TABLE 2 Behavioral validity tests of the SOEP questions, HL task, and IG.

			SOEP	HL	IG
Experimental tasks	Dohmen et al. (2011)	Real-stakes lottery experiment	☑		
	Vieider et al. (2013)	Real-stakes lottery experiment	☑		
	Lönnqvist et al. (2015)	Trust game	☑	☐	
	Charness et al. (2020)	Portfolio task	☑	☑	☑
		Mortgage task	☑	☑	☑
		Insurance task	☐	☑	☐
Stated real-world behavior	Dohmen et al. (2011)	Holding stocks	☑		
		Being self-employed	☑		
		Participating in sports	☑		
		Smoking	☑		
	Guenther et al. (2021)	COVID-19-related risky behavior	☐		
	Yang et al. (2022)	Smoking	☐		
		Taking antibiotics	☐		
		Drinking	☐		
	Verschoor et al. (2016)	Purchase of fertilizer			☑
		Growing cash corps			☐
		Proportion of sold agricultural output			☐
	Holden and Tilahun (2022)	Real-world investment behavior			☐
	Szrek et al. (2012)	Smoking	☑		☐
		Heavy drinking	☑		☐
		Not using the seat belt	☑		☐
		Engaging in risky sex	☑		☐
	Lusk and Coble (2005)	Consumption of genetically modified food		☑	
	Galizzi and Miniaci (2016)	Body mass index	☐	☑	
		Consumption of fruits and vegetables	☐	☑	
		Consumption of junk food	☐	☐	
		Smoking	☐	☐	
		Heavy drinking	☐	☐	
		Savings	☐	☐	
		Savings horizon	☐	☐	
		Saving regularly	☐	☐	
		Personal pension	☐	☐	

TABLE 2 (Continued)

		SOEP	HL	IG
Charness et al. (2020)	Savings amount	☐	☐	☐
	Share of risky investments	☐	☐	☐
	Owning real estate	☐	☐	☐
	Insurance demand	☐	☐	☐
	Being self-employed	☐	☐	☐

Note: This table provides an overview of studies that tested the behavioral validity of the SOEP risk questions, the HL task, and the IG. The box (☐ or ☑) indicates that the study tested the behavioral validity of the corresponding risk attitude measure. Only if the box is ticked (☑), the risk attitude measure can explain the corresponding behavior in the experimental task/stated real-world behavior.

Abbreviations: COVID-19, coronavirus disease 2019; HL, Holt and Laury; IG, Investment Game; SOEP, Socioeconomic Panel.

best predictor for holding stocks (Dohmen et al., 2011). However, when it comes to the overall explanatory power in the different contexts, the GRQ performs best (Dohmen et al., 2011). Guenther et al. (2021) and Yang et al. (2022) cannot confirm Dohmen et al.'s (2011) findings.

In Guenther et al.'s (2021) study, the SOEP questions are not able to explain risky behavior in the context of the COVID-19 crisis (leaving the house, wearing a mask, having contact to others inside and outside, etc.). In Yang et al.'s (2022) data, the GRQ cannot predict health-related behavior (such as smoking, taking antibiotics, or drinking). The evidence regarding the IG and the HL task is similarly mixed. While Verschoor et al. (2016) find that IG choices are associated with the purchase of fertilizer in their sample of farmers from rural Uganda, they find no relation with other behaviors involving risk (e.g., growing cash crops, etc.). Holden and Tilahun (2022) find that IG is not correlated with real-world investment behavior. Szrek et al. (2012) examine the predictive power of four risk attitude elicitation tasks, including the GRQ and the IG. They find that smoking, heavy drinking, not using a seat belt, and engaging in risky sex can be predicted by the GRQ but not by the IG. Regarding the HL task, Lusk and Coble (2005) find that it is significantly related to the consumption of genetically modified food. Galizzi and Miniaci (2016) find that the HL task is associated with the subject's body mass index, and the consumption of fruits and vegetables while the SOEP questions are not. Other behaviors involving risk (such as the consumption of junk food or the likelihood of saving regularly) cannot be predicted by either of the measures (SOEP questions and the HL task) (Galizzi & Miniaci, 2016). Charness et al. (2020) find that none of the three tasks is able to explain savings amount, share of risky investments, owning real estate, insurance demand, or being self-employed.

Again, the evidence on the behavioral validity with respect to stated real-life behaviors is rather mixed: While all measures are associated with some behaviors, they are not with others. However, no clear pattern is observable. For example, the HL task and the IG—that have a financial context—cannot reliably explain behaviors in financial contexts. On the other hand, some behaviors without financial context are associated with the HL task (e.g., the consumption of fruits and vegetables), but not with the IG. One potential explanation is that the laboratory setting of the experimental task may have biased the decisions (Charness et al., 2020) and the statements of real-world behavior might be biased through self-assessment difficulties or deliberate misstatements (e.g., due to social desirability) (Aklin et al., 2005; Yang et al., 2022).

2.3 | Correlations between the GRQ, HL, and the IG

In the following, findings on the correlations of the three measures that other studies have produced will be summarized. Table 3 provides an overview of the studies and the observed correlations—also mentions further measures that were examined in some of the studies, but they will not be considered in more detail.

Frey et al. (2017) tackle the question of whether risk attitudes are stable psychological traits by investigating 39 risk-taking measures.[4] They report that correlations between behavioral measures and self-report measures are generally rather low (mode of $\rho = 0.06$). Further, they find that while the correlations among the group of behavioral measures are weak (mode of $\rho = 0.08$), higher correlations are found among different self-report measures (mode of $\rho = 0.20$). The researchers interpret the results as an indication that self-report measures capture components of risk attitudes that are related, while behavioral measures capture different components that are unrelated to the ones of self-report measures and other behavioral measures.

Attanasi et al. (2018) compare three elicitation methods, namely, the HL task, the GRQ, and the lottery-panel task by Sabater-Grande and Georgantzis (2002).[5] In the HL task, subjects had to make 19 instead of 10 choices, thereby lottery A either paid €12 or €10, and lottery B either €22 or €0.50 (Attanasi et al., 2018). One of the two behavioral tasks (HL task and the lottery-panel task by Sabater-Grande and Georgantzis,(2002) was randomly selected to determine the payoff (Attanasi et al., 2018). The researchers find a relatively high and significant Spearman's rank correlation between the HL task and the GRQ ($\rho = 0.47$***). Further, they find, that the correlation is much higher for females ($\rho = 0.61$***) than for males ($\rho = 0.31$*) and for subjects who did not participate in a game theory class ($\rho = 0.52$***) than for subjects who participated ($\rho = 0.43$**). Finally, they also find that the correlation of the GRQ and the HL task is highest for subjects exhibiting a behavioral pattern of constant relative risk aversion in the lottery panel task by Sabater-Grande and Georgantzis (2002) ($\rho = 0.80$***).

Crosetto and Filippin (2013a) conduct a theoretical and empirical comparison of five behavioral tasks, among them the HL task and the IG.[6] Each subject completed only one of the behavioral tasks but all subjects answered the GRQ (Crosetto & Filippin, 2013a). The HL task was played with stakes twice as high as in the original task ($L_A = (€4.00, \quad p; \quad €3.20, 1-p)$, $L_B = (€7.70, \quad p; \quad €0.20, 1-p)$) and subjects were compensated by playing the preferred lottery in one randomly selected row (Crosetto & Filippin, 2013a). In the IG, subjects received an endowment of €4, and the investment was multiplied by 2.5 with a probability of 50% (Crosetto & Filippin, 2013a). While the researchers find that the HL task is significantly positively correlated with the GRQ ($\rho = 0.23$*), the relation of the GRQ and the IG is not significant ($\rho = 0.13$). In contrast to Attanasi et al. (2018), Crosetto and Filippin (2013a) note that the correlations are more pronounced for males than for females.

Galizzi and Miniaci (2016) investigate the test–retest stability, the behavioral validity, and the association of three elicitiation measures, including the HL task and three SOEP questions (the GRQ, and the financial- and health-context questions).[7] Thereby, subjects completed two

[4]Frey et al. (2017) conducted a within-subject design in Germany and Switzerland.
[5]Attanasi et al. (2018) employed a within-subject design in Italy.
[6]Crosetto and Filippin (2013a) conducted a between-subject design with a German sample.
[7]Galizzi and Miniaci (2016) conducted a within-subject design in a UK sample.

TABLE 3 Correlations of the SOEP questions, HL task, and IG in different studies.

Study	Sample	Elicitation measures	Correlation measure	Correlation coefficient	
Attanasi et al. (2018)	62 students in Milan, Italy	GRQ, HL, SG	Spearman	GRQ and HL	0.47***
Crosetto and Filippin (2013a)	444 (mainly) students in Jena, Germany	GRQ, HL, BART, IG, EG, BRET	Not reported	GRQ and HL	0.23*
				GRQ and IG	0.13
Frey et al. (2017)	1507 subjects in Basel, Switzerland and Berlin, Germany	39 behavioral and self-report measures	Pearson	Behavioral and self-report: 0.06 (significance levels not reported)	
Galizzi and Miniaci (2016)	661 (453) subjects in the first (second) wave, representative of UK population	GRQ, SOEP-f, SOEP-h, HL-h, HL-l, EG	Pearson	GRQ and HL-l	0.06/0.06
				SOEP-f and HL-l	0.19/0.04
				SOEP-h and HL-l	−0.02/−0.06
				GRQ and HL-h	0.01**/−0.06
				SOEP-f and HL-h	0.08/−0.0
				SOEP-h and HL-h	0.00/−0.13*
Lönnqvist et al. (2015)	232 subjects in Bonn, Germany	SOEP, HL	Spearman	SOEP and HL	−0.04
Menkhoff and Sakha (2017)	760 rural households in Thailand	GRQ, SOEP-f, IG, CE, EG, IQ	Spearman	GRQ and IG	0.03
				SOEP-f and IG	0.05
Szrek et al. (2012)	351 clients of health centers in South Africa	GRQ, HL, BART, DOSPERT	Spearman	GRQ and HL	0.02

Note: The table presents the correlations different studies have found of the SOEP questions with the HL task and the IG. The column "Elicitation measures" contains the abbreviations of measures used to investigate risk attitudes in the different data sets. For the sake of completeness, also measures not considered in this thesis are mentioned. The abbreviations relate to the following tasks: BART, balloon analog risk task by Lejuez et al. (2003); BRET, bomb risk elicitation task by Crosetto and Filippin (2013b); CE, certainty equivalent task used in Abdellaoui et al. (2011); DOSPERT, domain-specific risk-taking scale by Weber et al. (2002); EG, lottery-choice task by Eckel and Grossman (2002); GRQ, general risk question; HL (-l/-h), Holt and Laury lottery-choice task (with low/high stakes); IG, investment game by Gneezy and Potters (1997); SG, lottery-panel task by Sabater-Grande and Georgantzis (2002); IQ, hypothetical investment question following Barsky et al. (1997); SOEP (-f/-h), SOEP risk questions (with financial-/health-context). The Column "Correlation measure" indicates whether the correlation coefficient is a Spearman's rank correlation or a Pearson's correlation coefficient. The column "Correlation coefficient" contains the correlation coefficients for the measures considered in this thesis (see asterisks below). For the study of Galizzi and Miniaci (2016) two correlation coefficients are mentioned for each combination, as data were collected in two waves.

Abbreviations: HL, Holt and Laury; IG, Investment Game; SOEP, Socioeconomic Panel.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

HL tasks with different stakes. In the HL task with low stakes, lottery A paid either £40 or £32, and lottery B paid £77 or £2, and in the task with high stakes lottery A paid either £100 or £40, and lottery B paid £180 or £2 (Galizzi & Miniaci, 2016). Only 10% of participants received a task-dependent payoff, and this payoff was determined by one of the 91 answers subjects gave (Galizzi & Miniaci, 2016). The researchers collected data in two waves with a time lapse of 1 year. Regardless of the wave and the stakes of the HL task, they find that all Pearson's correlation coefficients between the HL tasks and the SOEP questions are extremely low, most are not significant, and some are even negative (correlation coefficients ranged from $\rho = -0.13^*$ to 0.19). Only the GRQ and the high-stakes HL task in the first wave as well as the SOEP health-context risk question and the high-stakes HL task in the second wave are significantly correlated ($\rho = 0.01^{**}$ and $\rho = -0.13^*$, respectively) (Galizzi & Miniaci, 2016).

Lönnqvist et al. (2015) investigate the correlation, the temporal stability, and the behavioral validity of the HL task and the seven SOEP questions.[8] The HL task was played with the original stakes, but in Euros instead of US dollars ($L_A = (€2.00, \quad p; €1.60, 1-p)$, $L_B = (€3.85, \quad p; €0.10, 1-p)$) and subjects were compensated by playing the preferred lottery in one randomly selected row (Lönnqvist et al., 2015). The estimated Spearman's rank correlation coefficient between the HL task and a factor of the seven SOEP questions is low and not significant ($\rho = -0.04$) (Lönnqvist et al., 2015).

Menkhoff and Sakha (2017) investigate the explanatory power of the IG, the GRQ, the SOEP financial-context risk question, and four other elicitation methods, for 11 risky behaviors.[9] In the IG, subjects received an endowment of 100 Thai Baht (approximately €2 at that time), and the investment tripled with a probability of 50% (Menkhoff & Sakha, 2017). The final payoff was determined by a random selection of one of the four incentivized tasks (Menkhoff & Sakha, 2017). The researchers find only low and nonsignificant correlations of the IG with the GRQ ($\rho = 0.03$) and SOEP financial question ($\rho = 0.05$).

Szrek et al. (2012) explore the predictive power for health-related risky behaviors of four elicitation measures, among them the GRQ and the HL task.[10] The GRQ scale ranged from 1 to 7, in contrast to the original SOEP scale (Szrek et al., 2012). The HL task was played with South African rand (R7.5 were $1 at that time), and lottery A paid either R25 or R20 and lottery B either R48 or R2 (Szrek et al., 2012). Subjects were compensated by playing the preferred lottery in one randomly selected row (Szrek et al., 2012). The researchers compute Spearman's rank correlations between the measures, and find that the correlation between the HL task and the GRQ is near zero and not significant ($\rho = 0.02$).

Summarizing these findings, two main observations can be made: First, the observed correlations vary greatly. For example, between the GRQ and the HL task significant correlations range from −0.13 to +0.47. Second, all correlation coefficients are below 0.5, and in many cases not significant. These observations put the reliability of the measures into question. Different explanations have been proposed: Some researchers suggest that low correlations may be the result of context-specific risk preferences (Reynaud & Couture, 2012; Weber et al., 2002), hence elicited risk attitudes are only indicative of risk preferences in the corresponding context. While Reynaud and Couture (2012) find evidence for this explanation,[11] Dohmen et al. (2011)

[8]Lönnqvist et al. (2015) employed a within-subject design with a German sample.
[9]Menkhoff and Sakha (2017) conducted a within-subject design with rural households in Thailand.
[10]Szrek et al. (2012) conducted a within-subject design in a sample of South African health center clients.
[11]Reynaud and Couture (2012) find that the HL task and the lottery-choice task by Eckel and Grossman (2002) are only significantly correlated with some contexts of a risk-taking psychometric questionnaire (the domain-specific risk-taking scale by Weber et al., 2002), namely, the financial and recreational context.

find that the correlations between the context-specific SOEP questions are large ($\rho > 0.4$). As already mentioned in Section 2.1, they also find that the GRQ is highly correlated with the context-specific SOEP questions and interpret their findings as an indication for the existence of an underlying risk preference trait. One would therefore expect that the GRQ is also significantly and positively correlated with the HL task and the IG have a financial context. However, this is only the case in the study of Attanasi et al. (2018) and for the HL task (but not the IG) in the study of Crosetto and Filippin (2013a). Context dependence of risk attitudes is also not an explanation in the considered studies, as the correlations between the SOEP financial-context risk question and the HL task/the IG are not significant in the study of Galizzi and Miniaci (2016) and Menkhoff and Sakha (2017).

Another explanation is that subjects have difficulties revealing their true risk preferences when there are comprehension problems with a task (Anderson & Mellor, 2009; Reynaud & Couture, 2012). As already mentioned in the previous section, especially the HL task appears to be complex (for some subjects) and inconsistent answers are a recurrent finding (Andersen et al., 2006; Charness & Viceisza, 2011; Dave et al., 2010). In contrast, the IG and the SOEP questions seem to be easier to understand (Charness et al., 2013; Holt & Laury, 2014; Lönnqvist et al., 2015; Verschoor et al., 2016). Anderson and Mellor (2009) suggest that subject traits such as the comprehension may be a factor that influences how strong the measures are correlated. This may possibly explain why the correlations of the GRQ and the HL task are higher in the study of Attanasi et al. (2018) and Crosetto and Filippin (2013a), where the sample consisted mainly of students. What cannot be explained is why in the data set of Crosetto and Filippin (2013a) the HL task but not the IG is significantly correlated with the GRQ, although the latter should be easier to understand.

Other researchers suggest that low correlations may be the result of the experimental design, especially the stakes involved in the tasks, the number of tasks subjects have to complete or the incentive structure (Attanasi et al., 2018; Crosetto & Filippin, 2013a; Holt & Laury, 2002). For example, a closer look at the addressed studies reveals that correlations with the HL task are higher when moderate stakes are used, as in the study by Attanasi et al. (2018) and Crosetto and Filippin (2013a) (payoffs range up to €22 and €7.7, respectively). In contrast, the correlations are marginal and (or) not significant in the study of Lönnqvist et al. (2015), where the stakes of the HL task were rather low (up to €3.85), and in Galizzi and Miniaci (2016), where the stakes were rather high (up to £180). Although differences in correlations cannot automatically be attributed to the level of stakes, it may be worthwhile to examine the effect of the monetary stakes on the correlations between the elicitation measures in more detail.

Furthermore, the number of tasks subjects had to complete might have an impact on the elicited risk attitudes (Crosetto & Filippin, 2013a). On the one hand, the validity of elicited risk attitudes may be affected as completing multiple elicitation measures may be tiring and the attention might fade (e.g., in the study of Frey et al. (2017) subjects completed 39 tasks in a daylong session (Anderson & Mellor, 2009). On the other hand, Crosetto and Filippin (2013a) note that completing multiple incentivized tasks might also lead to hedging behavior in subjects that are not risk averse (e.g., take little risk in the first task, then take a lot of risk in the second task; Crosetto & Filippin, 2013a). This in turn may result in low correlations (Crosetto & Filippin, 2013a).

Another factor that might impact elicited risk attitudes is the incentive structure: In some studies, only one of the multiple incentivized tasks is randomly selected to determine the final payoff (as in Attanasi et al. (2018) and Menkhoff and Sakha (2017)), or only a small randomly

selected share of subjects receives a task-dependent payoff (as in Galizzi and Miniaci, 2016). This might lead subjects to perceive the involved monetary stakes only as hypothetical payoffs if the likelihood that they impact their final payoff is small. Holt and Laury (2002) found that real monetary payoffs are important, as subjects behaved differently when the task had hypothetical stakes. These factors may have affected the observed correlations, as impacts on the elicited risk attitudes consequently also affect the correlation of different measures.

In summary, the HL task has been investigated much more than the IG. To the best of the author's knowledge, the correlation of the IG with the GRQ (or other SOEP questions) has only been examined in two studies—in Crosetto and Filippin (2013a) and Menkhoff and Sakha (2017)—that did not find any significant correlation. However, given that previous studies have suggested that the IG is less complex than the HL task (Crosetto & Filippin, 2013a), one would expect elicited risk attitudes to be more in line with true preferences and therefore more likely to correlate with the GRQ. With the next chapter, this thesis will extend the evidence on the association of the GRQ with the IG with an empirical analysis that differs especially in terms of sample size, endowments, and IG settings.

3 | ASSOCIATION OF THE GRQ AND THE IG

3.1 | Theoretical considerations

Motivated by the fact that evidence on the relationship between the GRQ and the IG is limited and that the few studies that investigated the relation found no significant correlation (Crosetto & Filippin, 2013a; Menkhoff & Sakha, 2017), we now analyze how the GRQ is associated with the IG. The relation will be investigated empirically with three samples of students from the University of Hamburg that are larger (115, 325, and 420 subjects) than the data set used in Crosetto and Filippin (2013a), which consisted of only 86 subjects. This is expected to improve the statistical power of potential effects (Cohen, 1988). Another particularity of the investigation is that in one of the data sets, the parameters of the IG are defined so that the expected return of investing is not positive, as in most other IGs, but negative (henceforth, IGs with positive/negative expected returns are called positive/negative IGs). As already mentioned, this has implications on how risky shares are interpreted in terms of risk attitudes. If it turns out that negative IGs are valid risk attitude measures, this is particularly interesting in light of the inability of positive IGs to disentangle risk-neutral from risk-seeking individuals.

3.2 | Data sets and experimental setups

The data used to examine the relations proposed in the previous section stem from three studies conducted by the Institute for Risk Management and Insurance at the University of Hamburg. Although the studies have different research emphases, they have in common that subjects played different versions of the IG and answered the GRQ. For the purpose of this thesis, the focus lies mainly on data that were collected on a basic version of the IG and answers to the GRQ. For the sake of statistical power, the three data sets will be merged as one big data set and analyzed altogether before they are analyzed individually. The experimental setup and the data utilized for the analysis are explained in the following sections.

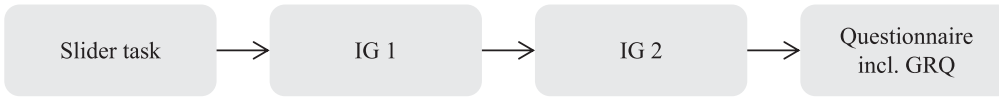

FIGURE 1 Experimental procedure of Data set 1. *Source*: Own illustration based on Hinck et al. (2022, p. 24).

3.2.1 | Data set 1: Impact of background risk on risk taking

The first data set stems from the study by Hinck et al. (2022), who investigated the impact of background risks on risk taking. The experiment was conducted in August 2021 with 325 subjects in the lab at the University of Hamburg.

The following description of the experimental setup is based on Hinck et al. (2022) and is illustrated in Figure 1: First, subjects completed a task, where sliders had to be aligned within a specified time, to "mitigate concerns of a potential house-money effect" (Hinck et al., 2022, p. 23). Next, subjects were confronted with two versions of the IG. One was similar to the IG presented in Section 2.3. The other one included a multiplicative background risk, where the payoff of the "standard" IG is exposed to an additional risk. The order in which subjects completed the two tasks was randomized. Additionally, subjects completed a questionnaire on their sociodemographic background, which included the GRQ. The final payoff was determined by a random selection of one of the two IGs.

For the purpose of this paper, only the choices made in the standard version of the IG (without background risk), answers to the GRQ, and sociodemographic data are investigated. Investment decisions in the IG with background risk are not included, as the effect of the background risk might bias the results. In the standard IG, subjects received an endowment of €8 ($=W_0$), and with a probability of 50% ($=p$), the asset yielded a return of 2.5 ($=k$); otherwise, the investment was lost. As a result, the IG had a positive expected return of investing ($2.5 \times 0.5 = 1.25 > 1$). As all subjects completed the GRQ and the standard IG, the data set contains 325 observations.

3.2.2 | Experiment 2: Impact of relative wealth placement on risk taking

The second data set was collected by Hillebrandt and Steinorth (2022) for a study investigating whether individuals behave differently in the IG, if they know the endowments of their peers and their own wealth standing within a group. The experiment took place from April to June 2019 with 420 subjects at the WiSo-Experimentallabor of the University of Hamburg.

Hillebrandt and Steinorth (2022) describe the experimental setup as follows (the steps are illustrated in Figure 2): After an introduction, subjects faced their first IG. Next, a cognitive reflection test was completed, measuring how prone subjects are to behavioral biases and overconfidence. Afterward, subjects played a second IG. They then completed a questionnaire testing their financial literacy. After that, a third IG was conducted. The participants then received their final payoff (determined by a random selection of one of the IGs) and completed a questionnaire on demographic information that included the GRQ. After that, another experiment was conducted.

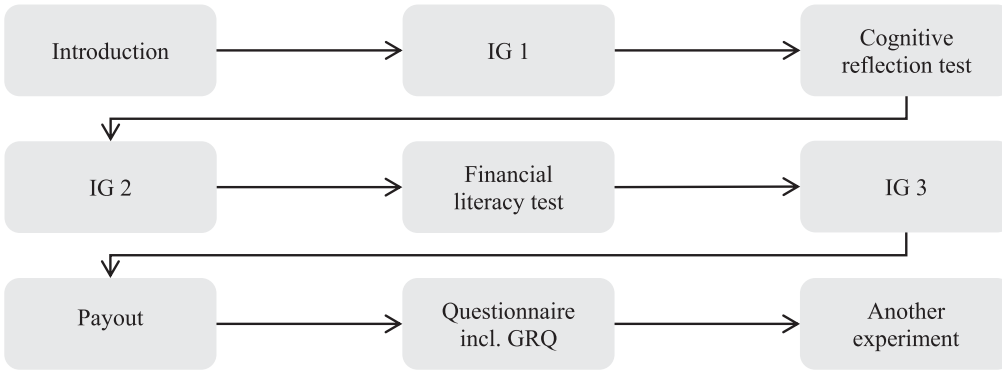

FIGURE 2 Experimental procedure of Data set 2. *Source*: Own illustration based Hillebrandt and Steinorth (2022, p. 29).

The IG endowments were presented in Taler, with 100 Taler corresponding to €0.25. Every participant played the IG twice in a social setting,[12] and once in a control setting that was similar to the standard IG version (without grouping). The order in which subjects played each setting was randomized. In this thesis, the focus lies on the data gathered in the control setting, as subjects' decisions were not influenced by the knowledge of the endowments of other players. In the control setting, subjects randomly either received 3450 (€8.63) or 4050 Taler (€10.13). All settings have in common that the positive development of the risky asset had a probability of 50% ($= p$) and multiplied the investment with 2.5 ($= k$). As a result, the expected return of investing was positive ($2.5 \times 0.5 = 1.25 > 1$).

Besides the data from the control setting of the IG and the GRQ, sociodemographic information is included in the analysis as controls. As all subjects took part in the control setting of the IG, the relevant data set contains 420 observations.

3.2.3 | Data set 3: Impact of a dark triad personality on gambling behavior

The third data set was collected by Jaeger and Steinorth (2022) for a study examining the impact of a dark triad personality on gambling behavior in the IG. The concept of the dark triad was first introduced by Paulhus and Williams (2002) and consists of three personality traits—Machiavellianism, narcissism, and psychopathy—that typically exhibit manipulative, self-centered, and callous behavior (Harrison et al., 2018). Jaeger and Steinorth (2022) investigate whether individuals who score high on the dark triad are more prone to excessive risk-taking. The experiment was conducted with 363 subjects at the WiSo-Forschungslabor at the University of Hamburg from July to August 2021.

Jaeger and Steinorth (2022) describe the experiment as follows (the steps are illustrated in Figure 3): First, subjects completed the HL task. After that, they were asked to fill out a survey on demographic information, which included the GRQ. Next, subjects played the IG. The last

[12]Subjects were grouped into three, all members receiving different endowments. Subjects were informed about their relative wealth standing in the group.

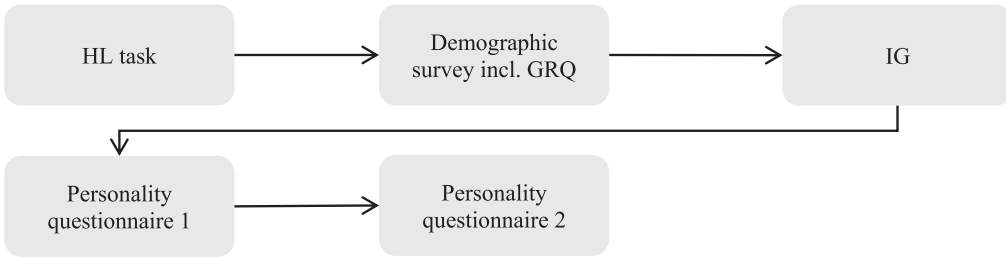

FIGURE 3 Experimental procedure of Data set 3. *Source*: Own illustration based on Jaeger and Steinorth (2022, p. 8).

part consisted of completing two personality questionnaires.[13] The final payoff was determined by the HL task, the IG, and attention checks that were incorporated into the questionnaires.

Different treatments were used in the IG to investigate whether dark triad personalities behave differently and results are robust to the impact of the relative wealth position as well as the skewness level of the risky investment. In sum, subjects were assigned to one of the six treatment groups that are presented in Table 4. Like in Data set 2, some of the IG treatments were played in a social setting[14] (low/high) and others in the standard IG setting without grouping (control). In the control treatment, subjects randomly either received €7 or €9 as initial endowment. Additionally, the treatments differed concerning the skewness, hence the parameters of the IG: Three treatments were played in a "default" setting with $p = 0.3$ and $k = 3$, and the others in a "skewed" setting with $p = 0.1$ and $k = 9$.[15] Note that in contrast to Data sets 1 and 2, both skewness settings in Data set 3 lead to a negative expected return of investing in the IG ($3 \times 0.3 = 0.9 < 1$ and $9 \times 0.1 = 0.9 < 1$).

We include subjects that played the IG in the control settings ("Default Control" and "Skewed Control"), where subjects were not influenced by social effects through a grouping. This results in 115 observations. Besides the data from the IG and the GRQ, demographic information is included in the analysis as controls.

3.3 | Results

3.3.1 | Summary statistics

Table 5 exhibits the summary statistics for a Merged Data set consisting of the three data sets, as well as of the individual data sets. The Merged Data set consists of 860 observations, 325 from Data set 1, 420 from Data set 2, and 115 from Data set 3; 62.4% of respondents are female in the Merged Data set. Also, in all three individual data sets there are more female than male

[13]The personality questionnaires were the Short Dark Triad questionnaire (developed by Jones & Paulhus, 2014), which aims to measure the extent of dark triad personality traits, and the hypersensitive narcissism questionnaire (developed by Hendin & Cheek, 1997), which aims to distinguish between overt and covert narcissists Jaeger and Steinorth (2022). The order was randomized.

[14]Subjects were grouped into three, all members receiving different endowments. Subjects were informed about their relative wealth placement in the group.

[15]The skewness was varied as literature has shown that "skewness is attractive in betting" (Jaeger & Steinorth, 2022, p. 3).

TABLE 4 Treatment overview of Data set 3.

Treatment	Number of group members	Initial endowments in € (W_0)	Probability (p) of positive outcome (%)	Multiple for positive outcome (k)
Default control	-	7/9	30	3
Default low	3	5/7/9	30	3
Default high	3	7/9/11	30	3
Skewed control	-	7/9	10	9
Skewed low	3	5/7/9	10	9
Skewed high	3	7/9/11	10	9

Note: This table is taken from Jaeger and Steinorth (2022, p. 11).

subjects (female share ranging from 60.7% to 64.3%). Hinck et al. (2022) suggest that the reason for this unequal gender distribution might be that the laboratory is located close to the humanities and social sciences faculties.

In the Merged Data set, subjects are on average 25.7 years old. The lowest average age can be observed in Data set 3 (24.8 years), while Data set 2 has the highest average age (25.9 years) and also the widest age range (16–81 years). In Data set 3, there is one observation where the age was not indicated. To avoid deleting the observation, the missing value was filled with the (rounded) average age of 25 years.

In all data sets, the majority of subjects have a high school diploma (46.5%–53.9%) or a bachelor's degree (31.3%–32.6%) as the highest level of obtained education. The share of subjects with an economics-related major range from 39.1% to 43.8%.

The average GRQ score in the Merged Data set, Data sets 1 and 2, is approximately 4.9. This finding is in line with previous studies: For example, average GRQ scores are 4.4 and 4.8 in Dohmen et al. (2011), and 4.5 in Galizzi and Miniaci (2016). In Data set 3, the average GRQ score is slightly higher, namely, 5.0. Only in Data set 2 (and consequently in the Merged Data set), the ranges picked in the GRQ spread across the whole scale (from 0 to 10). In all data sets, there are more subjects that picked a GRQ score lower (42.6%–49.8%) than higher (37.1%–40.0%) than 5.

In the Merged Data set, 42.4% of the endowment is invested in the IG on average. The highest average risky share is found in Data set 1 (52.8%) and the lowest in Data set 2 (35.6%). While the studies that introduced the IG—Gneezy and Potters (1997) and Charness and Gneezy (2010)—found higher average shares (44.8%–71.9% and 70.6%, respectively), the findings of subsequent studies are more in line with the present results: The average share invested in the IG is 36.4% in Menkhoff and Sakha (2017), 44% and 43% in Holden and Tilahun (2022), 46.5% in Charness et al. (2020), and 47.8% in Charness and Viceisza (2011). In all data sets, the risky shares spread across the whole range (some subjects invest nothing, others invest everything). Recall that in Data set 3, the expected return of investing is negative; hence, only risk-seeking individuals should invest. Surprisingly, 93.9% of individuals invest a positive amount in the IG.

3.3.2 | Association of GRQ scores and risky shares in the IG

To test the relation of the GRQ and the IG statistically, first Spearman's rank correlation coefficient is calculated. This allows to investigate whether the ordering of individuals

TABLE 5 Summary statistics.

	Merged Data set (Data sets 1–3)	Data set 1	Data set 2	Data set 3
Observations	860	325	420	115
Gender (%)				
Female/male	62.4/37.6	64.3/35.7	60.7/39.3	63.5/36.5
Age (years)				
Mean (SD)	25.7 (5.633)	25.7 (4.933)	25.9 (6.275)	24.8 (4.924)
Minimum	16	19	16	18
Maximum	81	58	81	55
Degree (%)				
High school	49.5	46.5	50.7	53.9
Bachelor	32.4	32.6	32.6	31.3
Other	18.1	20.9	16.7	14.8
Major (%)				
Economics related	41.7	40.0	43.8	39.1
Other	58.3	60.0	56.2	60.9
GRQ score				
Mean (SD)	4.90 (2.108)	4.93 (2.115)	4.85 (2.192)	5.03 (1.754)
Minimum	0	1	0	2
Maximum	10	10	10	9
Score </=/> 5 (%)	47.9/13.8/38.3	47.4/13.5/39.1	49.8/13.1/37.1	42.6/17.4/40.0
Risky share in IG (%)				
Mean (SD)	42.4 (28.917)	52.8 (27.977)	35.6 (28.338)	37.7 (24.963)
Minimum	0	0	0	0
Maximum	100	100	100	100
$\delta = 0/ > 0/ = 100$	5.8/83.1/11.1	4.3/79.1/16.6	6.9/84.5/8.6	6.1/89.5/4.4

Note: The table provides a descriptive overview of the data used for the empirical analysis. Standard deviations are reported within parentheses. For the GRQ score, the row "score 5 (%)" indicates the percentage of individuals who picked a sore lower, equal, or higher than 5 in the GRQ. For the risky share in the IG, the row "$= 0/ > 0/ = 100$" indicates the percentage of subjects who invested nothing/a share/the whole endowment in the IG.

Abbreviations: GRQ, General Risk Question; HL, Holt and Laury; IG, Investment Game.

according to their risk preferences is similar in the GRQ and in the IG. Table 6 presents the results and shows that the rank correlation is positive and highly significant in all data sets. In the Merged Data set, the rank correlation is $\rho = 0.36^{***}$, indicating that there is a positive interdependence between the GRQ and the IG; hence, risky IG shares tend to increase with the GRQ scores. With a correlation coefficient of $\rho = 0.46^{***}$, this relation is even more pronounced in Data set 1, while the correlation is lowest in Data set 3 ($\rho = 0.26^{***}$).

TABLE 6 Spearman's rank correlation of the GRQ score and the risky share in the IG.

	Merged Data set (Data sets 1–3)	Data set 1	Data set 2	Data set 3
Answer to GRQ	0.3638***	0.4581***	0.3543***	0.2577***

Abbreviations: GRQ, General Risk Question; IG, Investment Game.

***$p < 0.01$ indicate the level of significance.

Next, an ordinary least-squares (OLS) regression analysis is conducted to investigate the relation of the two measures in more detail.[16] With the risky IG share as the dependent variable and the GRQ score as the main independent variable, the regression tests the null hypothesis, "There is no relationship between the risky share in the IG and the score in the GRQ." To test the robustness of the results, a second model with controls is estimated for each data set. The regression results are reported in Table 7.

In the Merged Data set without controls (column 1), the GRQ coefficient is positive and significant. The risky share on average increases by 5 percentage points with every additional score point picked in the GRQ. The introduction of control variables does not alter this relation (column 2). The three individual data sets yield similar results, with the coefficients differing only slightly. The highest impact is found in Data set 1, where the risky share increases by 6 percentage points on average (columns 3 and 4), while the smallest impact is found in Data set 3 (4 percentage points) (columns 7 and 8). In sum, the results indicate that in all data sets the GRQ scores are able to predict the risky shares in the IG. Hence, the null hypothesis can be rejected.

The controls provide information about the plausibility of the results. In the Merged Data set, the gender coefficient is significant and indicates that the risky share is on average 6 percentage points lower for females than for males (column 2). This result is in line with previous findings that females are significantly more risk averse than males (e.g., see Dasgupta et al., 2016, 2019; Dohmen et al., 2011; Galizzi & Miniaci, 2016; Guenther et al., 2021). Gender also has a significant impact in the three individual data sets (columns 4, 6, and 8); however, the relation goes in the opposite direction in Data set 3 (females on average invest 11 percentage points more than males).

A significant impact that was found only in Data set 3 concerns the impact of age (column 8): With every additional year of age, the risky share increases by 1 percentage point on average. In contrast to this, Dohmen et al. (2005) find that subjects on average invest less with increasing age, in an investment task that resembles the IG of Gneezy and Potters (1997).

Different initial endowments were used in the three data sets (Data set 1: €8; Data set 2: €8.63 and €10.13 (converted in Taler); Data set 2: €7 or €9). The regression results indicate that the endowment has a significant impact (at the 10% significance level) in Data set 3, indicating that the risky share decreases with increasing endowment (column 8). This result is surprising, as the endowments differed only by €2. In contrast, in the Merged Data set and Data set 2, the endowment has no significant impact on the risky share (columns 2 and 6). Another interesting finding concerns the binary variable "currency" in the Merged Data set (column 2): To investigate the effect of the currency (Euros or Taler) on the risky share, the two data sets that used Euros (Data sets 1 and 3) were attributed with the value 1, and Data set 2, which used

[16]The data requirements for the OLS regression have been checked; details can be found in Appendix B.

TABLE 7 OLS regression results on the risky share in the IG.

	Dependent variable: Share invested in the IG							
	Merged Data set (Data sets 1–3)		**Data set 1**		**Data set 2**		**Data set 3**	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
GRQ	0.05*** (0.005)	0.05*** (0.005)	0.06*** (0.007)	0.06*** (0.007)	0.05*** (0.007)	0.04*** (0.007)	0.04*** (0.013)	0.04*** (0.013)
Gender	-	−0.06*** (0.020)	-	−0.09*** (0.032)	-	−0.09*** (0.029)	-	0.11** (0.046)
Age	-	0.001 (0.002)	-	0.0003 (0.002)	-	0.0001 (0.002)	-	0.01** (0.005)
Economics-related major	-	0.001 (0.018)	-	−0.01 (0.028)	-	0.003 (0.026)	-	0.005 (0.043)
Endowment	-	−0.01 (0.014)	-	-	-	0.01 (0.017)	-	−0.04* (0.022)
Currency	-	0.15*** (0.026)	-	-	-	-	-	-
IG setting	-	0.15*** (0.034)	-	-	-	-	-	-
Skewed	-	−0.005 (0.046)	-	-	-	-	-	−0.02 (0.042)
Constant	0.16*** (0.024)	0.07*** (0.137)	0.21*** (0.035)	0.29*** (0.079)	0.12*** (0.034)	0.14 (0.163)	0.18*** (0.063)	0.17 (0.277)
R^2	0.155	0.245	0.233	0.253	0.143	0.167	0.076	0.194
Observations	860	860	325	325	420	420	115	115

Note: This table shows the coefficients of OLS regression models for the Merged Data set (consisting of Data sets 1–3) and the three individual data sets. The risky IG share is the dependent variable and the GRQ score is the main independent variable. Due to heteroscedasticity, the regression parameters are calculated on robust standard errors (Hayes & Cai, 2007), which are presented within parentheses. "Gender" takes the value 1 for females and 0 for males. "Age" is scaled in years. "Economics-related major" is a dummy variable that takes the value 1 if individuals have an economics-related major, and 0 otherwise. "Endowment" is a metric variable that takes the value of the corresponding endowment in €. "Currency" is a dummy variable that takes the value 1 if the IG was played with € (like in Data sets 1 and 3) and 0 if it was played with Talers (Data set 2). "IG setting" is a dummy variable that takes the value 1 for positive and 0 for negative IGs. "Skewed" is a dummy variable that takes the value 1 for the skewed setting () in Data set 3, and 0 otherwise.

Abbreviations: GRQ, General Risk Question; IG, Investment Game; OLS, ordinary least-squares.

*$p < 0.10$, **$p < 0.05$, and ***$p < 0.01$ indicate the significance level.

Talers, was attributed the value 0. The currency coefficient is significant and indicates that 15 percentage points more are invested in the IG if the endowments are presented in Euros instead of Talers. A possible explanation for this finding may be that although subjects were aware of the exchange rate in Data set 2, the unfamiliarity with the fictitious currency might have led to reluctance in the investment behavior.

The Merged Data set shows that the setting of the IG, hence, whether the IG has a positive or negative expected return, it has a significant impact on the risky share (column 2). On average, 11 percentage points more are invested in positive IGs (Data sets 1 and 2) than in the negative IG (Data set 3). This finding is intuitive, as in positive IGs even risk-averse individuals might invest a share of the endowment, while only risk-seeking individuals should invest "something" in negative IGs. In other words, as risk-neutral and risk-averse individuals should invest nothing in negative IGs, this should lead to a lower average risky share.

There are two controls that have no significant impact on the risky share in any of the data sets: Whether subjects have an economics-related major and the skewness of the IG setting.

The explanatory power in the Merged Data set is reasonable ($R^2 = 0.155$ without and $R^2 = 0.245$ with controls). For Data set 1, the explanatory power is even higher with $R^2 = 0.233$ in the model without and $R^2 = 0.253$ in the model with controls. It stands out that the explanatory power of Data set 3 without controls is relatively low ($R^2 = 0.076$) and increases substantially when controls are added ($R^2 = 0.194$). In contrast, in Data sets 1 and 2, the explanatory power in the models without controls is much higher, and increases only slightly when the controls are added. Consequently, the GRQ variable appears to be responsible for a relatively large part of the overall explanatory power in Data sets 1 and 2, while in Data set 3, the controls have a larger impact on the risky share.

The results can be summarized follows: In all data sets, the correlations between the GRQ scores and the IG shares are significant and positive, and the regression results reflect that the GRQ score can predict the risky share in the IG. This result is robust to added controls and a binary transformation of the GRQ. Although the relation is less pronounced in Data set 3, where the IG had a negative expected return (lower correlation and lower explanatory power of the GRQ), in sum, the results are evidence in support of Hypothesis 1 and indicate that answers to these two measures of risk attitude are positively associated.

4 | CONCLUSION

In the last three decades, the elicitation of risk attitudes has attracted a great deal of research. One reason for this is that after the development of measures boomed in the 1990s and 2000s, subsequent studies often found that the relations between different measures were only weak. This thesis delved into the discussion on the association of risk attitude measures by exploring three elicitation methods, namely, the risk questions introduced in the German SOEP including the GRQ, the HL task developed by Holt and Laury (2002), and the IG by Gneezy and Potters (1997).

After reviewing the strengths and weaknesses of the measures as well as existing evidence on their association, we provide new insights on the association between IG and GRQ, as evidence on this relation is limited to date. Three data sets are used for the empirical analysis, with the specialty that two different versions of the IG are included. The IG versions differ with respect to whether the expected return of investing is positive or negative, which has implications on the expected behavior.

With respect to the research question—how the GRQ is associated with the two IG versions—the empirical analysis revealed the following main patterns: First, in all data sets (and hence, in both IG versions), there is a reasonable significant and positive association between risky shares subjects choose in the IG and their GRQ score (correlations range from 0.26 to 0.46). Considering that previous studies were unanimous in finding no significant association (see Crosetto and Filippin (2013a) and Menkhoff and Sakha (2017), this finding is rather surprising and strengthens the legitimacy and relevance of the IG.

Note that the subject from the experimental lab is not representative of the German population in several dimensions: Subjects are younger, have better education, and are more likely female. The results of this study encourage further investigation into the circumstances under which the GRQ and IG are significantly associated.

ACKNOWLEDGMENTS

We gratefully acknowledge the work of Marc-Andre Hillebrandt, Sebastian Hinck, and Tim Jaeger in programming and conducting the three experimental studies we utilized in our metastudy. The second and third experiments were funded by the Deutsche Forschungsgemeinschaft (DFG) under grant number 433254283. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Petra Steinorth http://orcid.org/0000-0001-8728-2861

REFERENCES

Abdellaoui, M., Baillon, A., Placido, L., & Wakker, P. P. (2011). The rich domain of uncertainty: Source functions and their experimental implementation. *American Economic Review*, *101*(2), 695–723.

Aklin, W. M., Lejuez, C. W., Zvolensky, M. J., Kahler, C. W., & Gwadz, M. (2005). Evaluation of behavioral measures of risk taking propensity with inner city adolescents. *Behaviour Research and Therapy*, *43*(2), 215–228.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, *9*(4), 383–405.

Anderson, L. R., & Mellor, J. M. (2009). Are risk preferences stable? Comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty*, *39*(2), 137–160.

Attanasi, G., Georgantzís, N., Rotondi, V., & Vigani, D. (2018). Lottery- and survey-based risk attitudes linked through a multichoice elicitation task. *Theory and Decision*, *84*(3), 341–372.

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2018). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Springer Gabler.

Barsky, R. B., Juster, F. T., Kimball, M. S., & Shapiro, M. D. (1997). Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *The Quarterly Journal of Economics*, *112*(2), 537–579.

Charness, G., Garcia, T., Offerman, T., & Villeval, M. C. (2020). Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? *Journal of Risk and Uncertainty*, *60*, 99–123.

Charness, G., & Gneezy, U. (2010). Portfolio choice and risk attitudes: An experiment. *Economic Inquiry*, *48*(1), 133–146.

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

Charness, G., & Viceisza, A. (2011). *Comprehension and risk elicitation in the field evidence from rural Senegal*. International Food Policy Research Institute.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Crosetto, P., & Filippin, A. (2013a). *A theoretical and experimental appraisal of five risk elicitation methods*. DIW.

Crosetto, P., & Filippin, A. (2013b). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, *47*(1), 31–65.

Dasgupta, U., Mani, S., Sharma, S., & Singhal, S. (2016). *Eliciting risk preferences: Firefighting in the field* (Working paper). United Nations University. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2742553

Dasgupta, U., Mani, S., Sharma, S., & Singhal, S. (2019). Internal and external validity: Comparing two simple risk elicitation tasks. *Journal of Behavioral and Experimental Economics*, *81*, 39–46.

Dave, C., Eckel, C. C., Johnson, C. A., & Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, *41*(3), 219–243.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. (2005). *Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey*. Deutsches Institut für Wirtschaftsforschung (DIW).

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.

Drichoutis, A., & Vassilopoulos, A. (2021). Intertemporal stability of survey-based measures of risk and time preferences. *Journal of Economics & Management Strategy*, *30*(3), 655–683.

Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, *23*(4), 281–295.

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.

Galizzi, M., & Miniaci, R. (2016). *Temporal stability, cross-validity, and external validity of risk preferences measures: Experimental evidence from a UK representative sample* (Working paper). London School of Economics and Political Science. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2822613

Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, *112*(2), 631–645.

Guenther, B., Galizzi, M. M., & Sanders, J. G. (2021). Heterogeneity in risk-taking during the COVID-19 pandemic: Evidence from the UK lockdown. *Frontiers in Psychology*, *12*, 643653.

Harrison, A., Summers, J., & Mennecke, B. (2018). The effects of the dark triad on unethical behavior. *Journal of Business Ethics*, *153*(1), 53–77.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*(4), 709–722.

Hendin, H. M., & Cheek, J. M. (1997). Assessing hypersensitive narcissism: A reexamination of Murray's Narcism Scale. *Journal of Research in Personality*, *31*(4), 588–599.

Hillebrandt, M.-A., & Steinorth, P. (2022). *Relative wealth placement and risk-taking behavior* (Working paper). University of Hamburg. https://hdl.handle.net/11159/410632

Hinck, S., Peter, R., & Steinorth, P. (2022). *Multiplicative background risk and risk taking: Theoretical predictions and experimental evidence* (Working paper). University of Hamburg.

Holden, S. T., & Tilahun, M. (2022). *Can the risky investment game predict real-world investments?* (Working paper). Norwegian University of Life Sciences. https://www.econstor.eu/handle/10419/262027

Holt, C., & Laury, S. (2014). Assessment and estimation of risk preferences, 2014. In M. Machina & K. Viscusi (Eds.), *Handbook of the economics of risk and uncertainty* (Vol. 1). Elsevier B.V.

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

Jaeger, T., & Steinorth, P. (2022). *Dark triad and gambling in investment decisions* (Working paper). University of Hamburg. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4184399

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3): A brief measure of dark personality traits. *Assessment*, *21*(1), 28–41.

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.

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.

Lusk, J. L., & Coble, K. H. (2005). Risk perceptions, risk preference, and acceptance of risky food. *American Journal of Agricultural Economics*, *87*(2), 393–405.

Maier, J., & Rüger, M. (2010). *Measuring risk aversion model—Independently* (Working paper). University of Munich.

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.

Menkhoff, L., & Sakha, S. (2017). Estimating risky behavior with multiple-item risk measures. *Journal of Economic Psychology*, *59*, 59–86.

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563.

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, *1*, 803–809.

Reynaud, A., & Couture, S. (2012). Stability of risk preference measures: Results from a field experiment on French farmers. *Theory and Decision*, *73*(2), 203–221.

Sabater-Grande, G., & Georgantzis, N. (2002). Accounting for risk aversion in repeated prisoners' dilemma games: An experimental test. *Journal of Economic Behavior & Organization*, *48*(1), 37–50.

Smigierski, J. (n.d.). *Stata lineare Regression Voraus setzungen*. https://www.beratung-statistik.de/statistik-beratung-infos/stata-tutorial/stata-regression-voraussetzungen/

Szrek, H., Chao, L.-W., Ramlagan, S., & Peltzer, K. (2012). Predicting (un)healthy behavior: A comparison of risk-taking propensity measures. *Judgment and Decision Making*, *7*(6), 716–727.

Urban, D., & Mayerl, J. (2011). *Regressionsanalyse: Theorie, Technik und Anwendung*. VS Verlag für Sozialwissenschaften|Springer Fachmedien Wiesbaden GmbH.

Verschoor, A., D'Exelle, B., & Perez-Viana, B. (2016). Lab and life: Does risky choice behaviour observed in experiments reflect that in the real world? *Journal of Economic Behavior & Organization*, *128*, 134–148.

Vieider, F. M., Lefebvre, M., Bouchouicha, R., Chmura, T., Hakimov, R., Krawczyk, M., & Martinsson, P. (2015). Common components of risk and uncertainty attitudes across contexts and domains: Evidence from 30 countries. *Journal of the European Economic Association*, *13*(3), 421–452.

Weber, E. U., Blais, A., & 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.

Wolf, C., & Best, H. (2010). *Handbuch der sozialwissenschaftlichen Datenanalyse*. VS Verlag für Sozialwissenschaften.

Yang, M., L. S. J. Roope, J. Buchanan, A. E. Attema, P. M. Clarke, A. S. Walker and S. Wordsworth, 2022, Eliciting risk preferences that predict risky health behavior: A comparison of two approaches, *Health Economics 31*(5), 836–858.

How to cite this article: Gaertner, C., & Steinorth, P. (2023). On the correlation of self-reported and behavioral risk attitude measures: The case of the General Risk Question and the Investment Game following Gneezy and Potters (1997). *Risk Management and Insurance Review*, *26*, 367–392. https://doi.org/10.1111/rmir.12250

APPENDIX A: SCATTERPLOTS OF THE GRQ SCORES AND THE RISKY IG SHARES

Figure A1

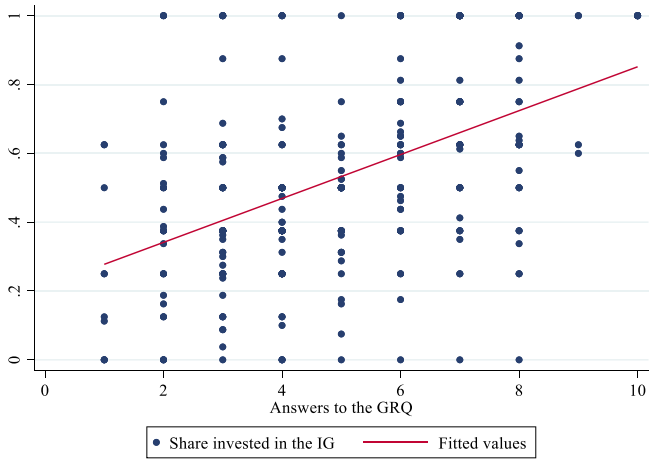

FIGURE A1 Scatterplot of the General Risk Question (GRQ) and Investment Game (IG) in Data set 1.

APPENDIX B: REGRESSION REQUIREMENTS AND RESULTS

VERIFICATION OF OLS REGRESSION DATA REQUIREMENTS

To ensure that OLS regression produces correct results, the data must fulfill certain requirements (Backhaus et al., 2018):

the relation between the dependent and the independent variables is linear,
the residuals have an expected value of zero,
there is no correlation between the independent variables and the residuals,
the residuals have a constant variance (homoscedasticity),
the residuals are not correlated (no autocorrelation),
there is no perfect multicollinearity between the independent variables,
the residuals are normally distributed.

MODEL WITH METRIC IG AND METRIC GRQ VARIABLE

Several tests have been conducted to test whether the data for the OLS regression with the risky shares in the IG as a dependent variable and the GRQ scores as the independent variable fulfill the requirements. The results are reported in Table B1 (requirements that are not met are highlighted by underlining).

Except for the homoscedasticity and normal distribution requirement (requirements (4) and (7)), all requirements are fulfilled. Regarding the homoscedasticity requirement, the tests show that the variance of the residuals is not constant in both models of the Merged Data sets and 2, and in Data set 3 in the model without controls. To avoid incorrect results in statistical tests resulting from heteroscedasticity, the regression parameters will be calculated on robust

TABLE B1 Verification of OLS regression requirements (metric IG, metric GRQ).

	Merged Data set (Data sets 1–3)		Data set 1		Data set 2		Data set 3	
Controls	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
(1) Linear relation	☑	☑	☑	☑	☑	☑	☑	☑
(2) Expected value residuals	~0	~0	~0	~0	~0	~0	~0	~0
(3) Correlation with residuals								
GRQ	0	0	0	0	0	0	0	0
Gender	-	0	-	0	-	0	-	0
Age	-	0	-	0	-	0	-	0
Economics-related major	-	0	-	0	-	0	-	0
Endowment	-	0	-	-	-	0	-	0
Currency	-	0	-	-	-	-	-	-
IG setting	-	0	-	-	-	-	-	-
Skewed	-	0	-	-	-	-	-	0
(4) Homoscedasticity	0.001	0.002	0.852	0.355	0.004	0.000	0.012	0.072
(5) Autocorrelation	1.865	2.045	2.192	2.188	1.979	1.960	1.938	1.845
(6) Multicollinearity	-	1.52	-	1.05	-	1.04	-	1.07
(7) Normal distributed residuals	0.000	0.000	0.015	0.439	0.000	0.000	0.062	0.062

Note: This table summarizes the results of tests that explore whether the requirements for the OLS are fulfilled. In row 1, the linear relationship of the dependent variable with the independent variables is assessed (requirement (1)). The relationships were examined in scatterplots, and the check sign (☑) indicates that an approximate linear relation was found between the dependent and independent variable(s). Row 2 shows the expected value of the residuals. For requirement (2) to hold, these values should be close to 0 (Backhaus et al., 2018). Row 3 exhibits the average correlation between the residual and the independent variables, which allows insights into whether the residuals are uncorrelated with the independent variables. The correlation should be 0 for requirement (3) to hold (Backhaus et al., 2018). Row 4 displays the *p* value of the Breusch–Pagan test for heteroscedasticity. A *p* value greater than 0.05 indicates that the variance of the residuals is constant (Smigierski, n. d.). In row 5, the results of the Durbin–Watson test are noted, which tests for autocorrelation between the residuals. As a rule of thumb, the Durbin–Watson statistic should be between 1.5 and 2.5 for requirement (5) to hold (Smigierski, n. d.; Urban & Mayerl, 2011). To check the independent variables for multicollinearity, the average VIF values are stated in row 6. These values should be less than 10 for requirement (6) to hold (Backhaus et al., 2018; Smigierski, n. d.). To test whether the residuals are normally distributed, the *p* values of the Shapiro–Wilk test are stated in row 7. Low *p* values indicate a significant deviation from the normal distribution (Wolf & Best, 2010); thus, the *p* values should be larger than 0.05 for requirement (7) to hold (Smigierski, n. d.). The underlined values do not fulfill the requirements.

Abbreviations: GRQ, General Risk Question; IG, Investment Game; OLS, ordinary least-squares.

standard errors (Hayes & Cai, 2007). Further, both models in the Merged Data sets and 2 as well as the model without controls in Data set 1 do not fulfill requirement (7) that the residuals should be normally distributed. This may affect the quality of statistical tests; however, the central limit theorem ensures that for samples with more than 40 observations, the significance tests are correct irrespective of the distribution of the residuals (Backhaus et al., 2018). As all data sets analyzed have more than 40 observations, this does not pose a problem in these data

sets. In sum, all requirements can either be met, flaws can be corrected or are not relevant in this context. Hence, the results obtained by the OLS regressions allow to explore the hypothesis.

Figure B1

FIGURE B1 Scatterplot of the General Risk Question (GRQ) and Investment Game (IG) in Data set 3.

Figure B2

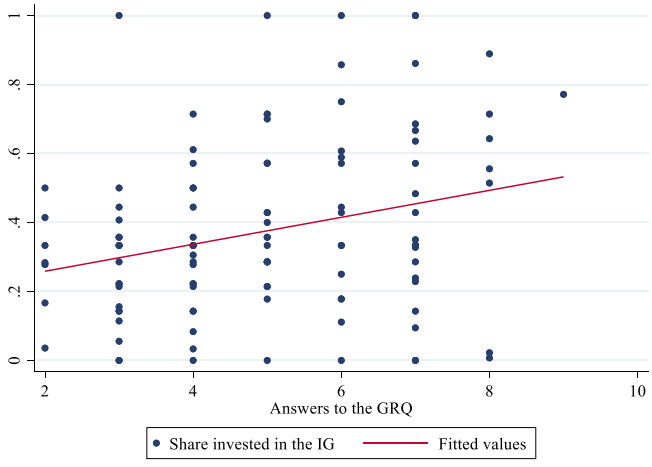

FIGURE B2 Scatterplot of the General Risk Question (GRQ) and Investment Game (IG) in Data set 2.