University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 6/2022 (382)

CALIBRATION AND INCENTIVES: EVIDENCE FROM CONTRACT BRIDGE

Michał Krawczyk
Maciej Wilamowski

Warsaw 2022

WORKING PAPERS 6/2022 (382)

UNIVERSITY OF WARSAW
FACULTY OF ECONOMIC SCIENCES
WORKING PAPERS

Calibration and incentives: evidence from contract bridge

Michał Krawczyk, Maciej Wilamowski*

University of Warsaw, Faculty of Economic Sciences
** Corresponding author: mwilamowski@wne.uw.edu.pl*

Abstract: We elicit probability forecasts from amateur contract bridge players. At the end of the auction of each deal in a tournament, the players were asked to make a guess (unobservable to others) about the probability with which the contract will be made. We observe them to be overall poorly calibrated. We also find that incentivizing correct forecasts makes no difference.

Keywords: overconfidence, forecast incentives, gender differences, bridge into professions

JEL codes: D01, L26, J24, J16

Introduction

As future is inherently uncertain, nearly all interesting predictions should in principle be made in probabilistic terms. Unfortunately, many studies report poor calibration of such predictions: it is not uncommon that in the class of cases in which the forecaster expects something to happen with probability *p*, in truth it happens with a systematically higher or lower probability. Making correct probabilistic forecast may be particularly difficult if the forecaster is not indifferent whether it happens or not: her optimism/wishful thinking or, by contrast, pessimism, may bias her judgment. Things get yet more complicated if the forecaster herself may actually *affect* the chance that something happens. Indeed, her overconfidence may lead her to overestimate the probability with which she will be able to make her desirable outcome come true (and conversely for underconfidence). It is perhaps not surprising, that making good probabilistic predictions about own performance may require long-time experience with given task, involving constructive feedback and perhaps direct incentives to predict correctly.

It should also be noted that this perhaps most demanding class of cases is of obvious importance. Indeed, it is often the person whose job is to *make something happen* that is also called upon to provide a prediction of *how likely* it is to happen. Even if these predictions are not said aloud, they may be vital for optimal decision making. For example, a basketball player must in each game make dozens of split-second assessments of her current probability of scoring – if it is too low, she should pass the ball instead.

It is thus highly desirable to deepen our understanding of the determinants of good probabilistic forecasts of own performance. One important dimension concerns incentives to predict correctly. On the one hand, they could encourage the forecaster to think hard and perhaps swallow her pride. On the other hand, if they are *too* strong compared to the stakes corresponding to the outcome itself, she may be tempted to give an unambitious prediction and then deliberately deliver correspondingly poor performance (a form of moral hazard).

Unfortunately, large majority of relevant studies involve un-incentivized predictions made in previously unfamiliar tasks performed in the lab: in a typical example, student participants are asked to guess what is the capital city of Brazil, what is the longest river in Europe etc. and then in each case report the probability with which they believe to be right. Whereas it may be interesting to observe that these numbers tend to be higher than the actual fraction of correct answers, it tells us little as to whether the result would change 1) if the same tasks were repeated; 2) if the tasks were more familiar and natural to the forecaster to begin with; 3) if there were incentives to predict correctly; and 4) depending on the outcome per se being desirable or not.

It is particularly surprising that only a small number of experiments addressed the natural question of incentives. In an early study, Shraw et al. (1993) found incentives to make correct probability forecasts to improve calibration. Yates, Lee, and Bush (1997) elicited two measures of confidence about correctness of answers to general knowledge questions. First, they were reported directly in response to a hypothetical question. Second, they were inferred from BDM-based evaluation of a gamble paying conditional on their answers being correct. Intriguingly, they found that their American sample, which was overconfident according to the hypothetical measure, became *more* overconfident in terms of their willingness to bet on own knowledge. By contrast, no impact was identified in the Chinese sample (which, in general, was even more overconfident).

Hoelzl and Rustichini (2005) asked their subjects to choose between ‘‘performance test’’ and ‘‘lottery’’. It was in subjects' best interest to opt for the former if and only if they believed they would perform above median in the test. Hoelzl and Rustichini (2005) found their subjects to be overconfident only when the task was easy. By contrast, when it was difficult, they switched to underconfidence, but only if success was rewarded with money. Given the design of the study, the results can be interpreted in terms of competitiveness, not so much self-confidence.

Grieco and Hogarth (2009) also used trivia tests in a laboratory, letting subjects choose between a random and performance-contingent payments. This time, unlike in the previous study, the choices were made after the task, making competitiveness dimension less silent. The main finding was that of over- (under)estimation showing up for hard (easy) tasks. Again there are some interpretational difficulties as discussed in Krawczyk (2012).

The latter paper reports field experiments in which subjects had to guess whether they are in the top half of the class on an exam. Female subjects more confident when incentivized to guess correctly.

The most closely related study is that by Keren (1987), the Baseline condition of our experiment being its replication. Because appreciating that paper requires understanding of the basic rules of contract bridge, we postpone the discussion of it until the next section.

The game of bridge

The rules of the game of bridge are relatively complex; for the purposes of this study it suffices to know the following. Bridge is played in pairs, with always two of them playing against each other. Each of the four players is dealt their own cards, so that they have their private information, inaccessible to others until it is too late. Each deal consists of two phases: *auction* and *play*. As a result of the auction one of the players becomes the *declarer* who is obliged to take at least a specific number of *tricks* during the play (aka "make the contract"). Her partner, called the *dummy*, is inactive during the play, with the declarer deciding how to play from both hands. In this sense, the declarer has the greatest control over the play and her skills are crucial for the success. Naturally, the two remaining players, the *defenders*, are generally trying to defeat the contract.

Players have several good reasons to think of how likely the declarer is to succeed (make the contract), particularly at the end of the auction and the start of the play. Indeed, the scoring rules are such that it typically pays to bid a contract if it is sufficiently likely to be made (but a better-scored contract is considerably less likely). The defenders may also be willing to use a *double*, which increases the penalty to the declarer in case she fails, if they believe this course of events is likely. Further, judged likelihood of the contract determines the optimal play. For example, it might make sense for the declarer to seriously fight for *over*tricks only if she believes the contract is unlikely to fail etc. Finally, when thinking of the best strategy of playing, players may come up with success forecasts as a by-product.

Still, these forecasts are rarely made in a very explicit way and never vocalized before the end of the play. It may also be noted that they may differ considerably between players in any specific deal, because of their private information, because of differences in their analytical skills, and because of their individual tendencies to be optimistic or self-confident (or not). The latter dimension is the focus of the current study.

Design and procedures

The field experiments were conducted during amateur bridge tournaments in Warsaw. The participants were told at the beginning of the tournament that they could participate in the study by making explicit probabilistic forecasts, at the end of the auction of each deal, concerning how likely it was that the declarer would make the contract, see Appendix A for the instructions and the cards used to elicit the forecasts.

We used two types of incentives. In the Baseline condition, three players from among those who have made at least some forecasts would be picked at random and receive 100 PLN each (ca. 23 euro). In the Incentivized condition, three such players would be picked at random and rewarded in accordance with the average Brier Score of their forecasts, with the perfect score resulting in the payment of 200 PLN and the lowest possible score resulting in the payment of zero. The participants were not given a very detailed description of the procedure in this case; they were just told that three randomly selected participants would be rewarded anything between 0 and 200 in accordance with accuracy of their forecasts. All the players in a given tournament were assigned to the same condition; else, the risk of treatment contamination would be very high.

In principle, a problem of moral hazard could occur, in that participants could be tempted to give very pessimistic predictions and play accordingly poorly to make them come true, reaching a high Brier Score. This is very unlikely though, for the following main reasons. First, participants are generally strongly motivated to do as well as possible in the tournament; second, playing poorly is not particularly enjoyable and the participants are there for fun, though this is not always apparent; third, such behaviour would very likely lead to a conflict with the partner; fourth, it could lead to being punished by the referee, earning a very bad reputation in the community etc. In any case, the experimental design allowed identifying such cases, should they ever occur. In particular, we would expect that such behaviour is more likely towards the end of the tournament and only in pairs that have been doing poorly so far.

We hypothesise that declarers will tend to make more positive forecasts than defenders. By investigating the forecasts made be the dummies, we will be able to know if the difference is due to optimism or overconfidence. Indeed, the contract being made is a positive outcome for the dummy, so optimism should affect their predictions, but cannot be affected by the dummy, so overconfidence should only affect the predictions of the declarers. If the dummies predict as highly as the declarers thus (higher than the defenders), optimism is at play; if the dummies predict as lows at the defenders, lower than the declarer, overconfidence is identified. We also hypothesise that incentives tend to reduce forecast errors.

Towards the end of the tournament the players were asked to fill in a short survey, see Appendix A.

Results

In total, we recorded 2318 predictions made by 190 players (157 males and 33 females) in four tournaments. On average, each player made a prediction in nearly 12 deals out of 27 played (see Table 1 for more details).

Table 1: The sample

incentive	**tournament**	**Gender**	**# players**	**# predictions**	**mean # predictions per player**
No	**17062019**	**F**	15	182	12.1
		M	47	564	12.0
	25042019	**F**	7	94	13.4
		M	44	586	13.3
Yes	**01072019**	**F**	8	80	10.0
		M	29	305	10.5
	16052019	**F**	3	68	22.7
		M	37	439	11.9

The distribution of predictions is very similar across roles (see Figure 1) with mean equal to 68.8 (67.6 for males and 74.2 for females).

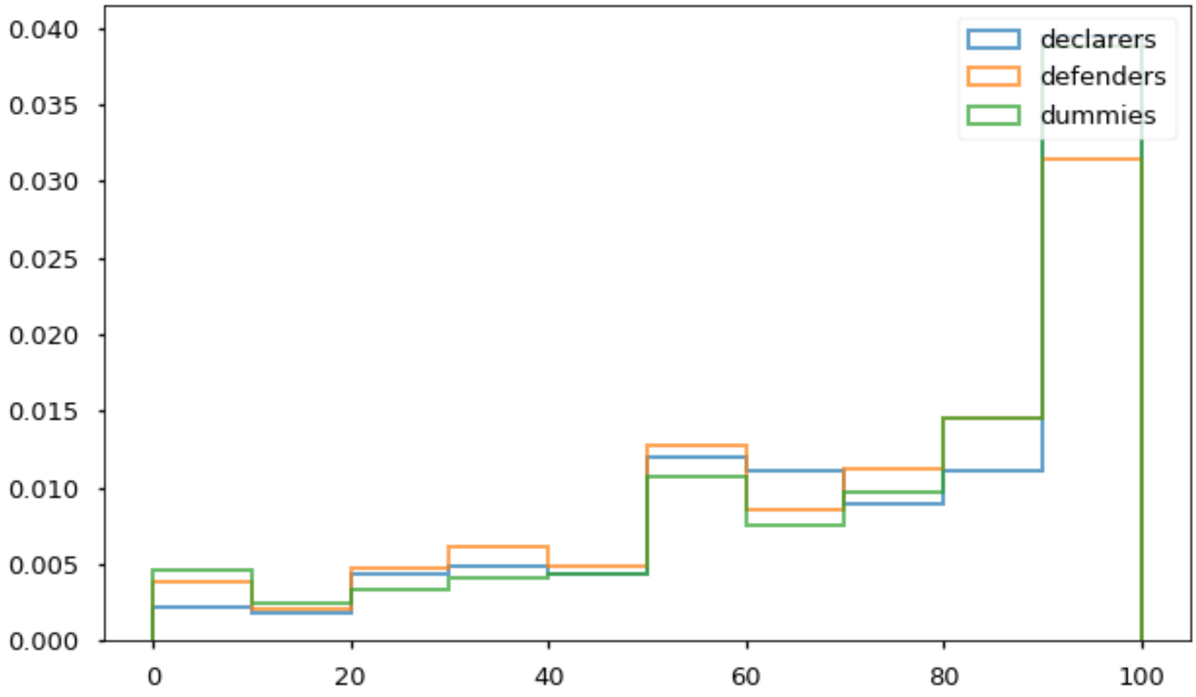

Figure 1: Distribution of predictions

Notes: (0=the contract will surely not be made to 100=the contract will surely be made)

The fraction of contracts actually made, by prediction bracket, is shown Figure 2. If players were perfectly calibrated, the dots would be aligned with the 45 degree line. For example, about 60% of contracts deemed to be made with probability 60% would actually be made and likewise for other levels of certainty. In fact, the slopes are positive but much smaller than one, with as many as about 40% of contracts deemed impossible or very unlikely being made and only about 75% of those deemed certain or almost certain. It is also interesting to note that the predictions made by players with different roles were very similar on average.

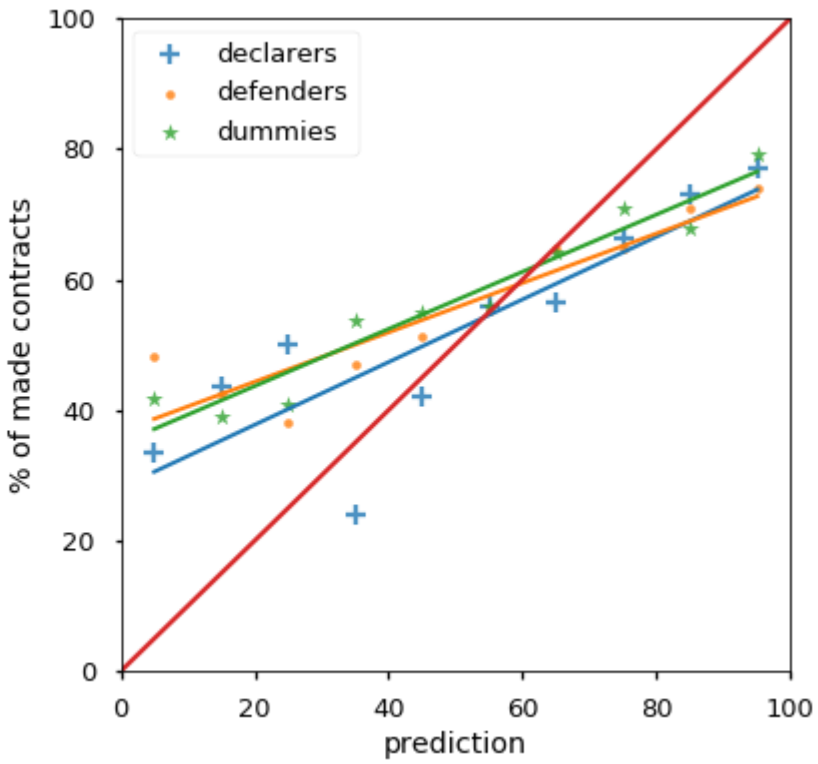

Figure 2. Share of contracts made, by player role and prediction bracket

Notes: We grouped predictions into 10 groups ([0-10), [10-20), …) and calculated the share of contracts made. The red line correspond to hypothetical, perfect calibration.

To perform econometric analyses, we calculated several measures. First, we calculated two measures of discrepancy between the prediction and the outcome:

Error = prediction - contract_made * 100

Absolute Error = |prediction - contract_made * 100|

As mentioned before, systematically high errors can be understood as overconfidence or optimism or overconfidence in declarers and as optimism in dummies.

Next, for each player in each tournament we calculated average values of above-mentioned measures to get Average Error and Average Absolute Error. The distribution of these four measures can be seen in Figure 3; all were used as dependent variables in econometric modelling.

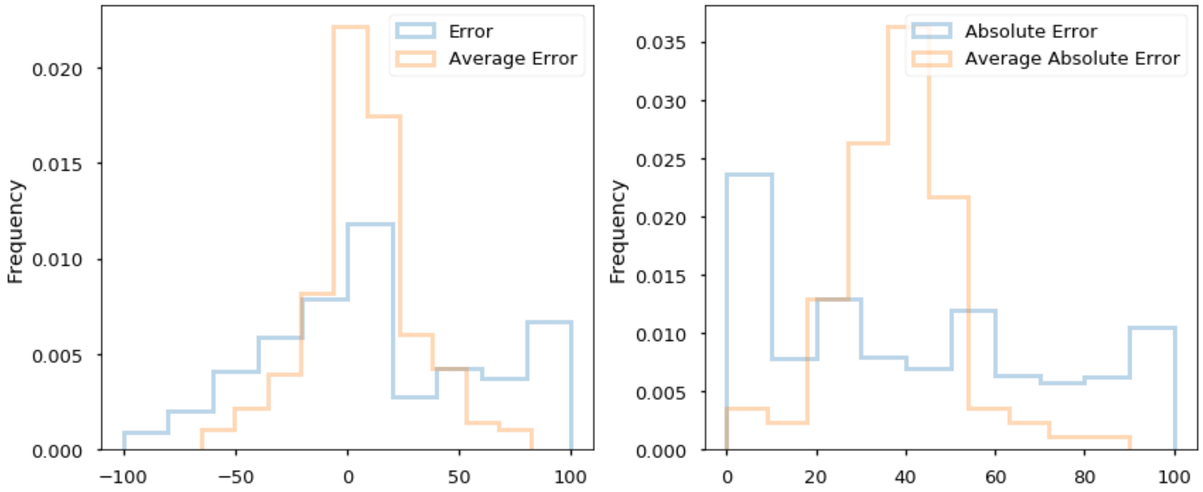

Fig. 3 Distribution of error measures.

As independent variables we used binary variables for gender, incentives and specific tournament (which had different hands dealt to players). Additionally, as a proxy for the difficulty of the contract, we calculated the average number of overtricks per player in each tournament.

In all the specifications we tested, we found only one significant variable, namely the number of overtricks. We did not find any support for relationship between target variables and other independent variables, including gender, role, and incentives, see Tables 2-5.

Table 2. Econometric estimation for Error

	1	**2**	**3**	4
Intercept	11.67*** (3.04)	11.22*** (2.43)	13.2*** (2.61)	11.29*** (1.91)
gender - M	-5.77* (2.64)	-6.89** (2.11)	-6.2** (2.12)	-6.92** (2.11)
role defender	1.31 (2.49)	-0.87 (1.99)	---	---
role declarer	3.89 (2.92)	3.47 (2.34)	---	---
incentive	-2.36 (2.1)	-0.85 (1.68)	---	---
number of overtricks	---	-19.45*** (0.54)	-19.47*** (0.54)	-19.43*** (0.54)

tournament 16052019	---	---	-5.34* (2.66)	---
tournament 17062019	---	---	-0.28 (2.46)	---
tournament 25042019	---	---	-4.25 (2.51)	---
R^2	0	0,36	0,36	0,36
R^2 adj.	0	0,36	0,36	0,36
Number of observations	2318	2318	2318	2318

Notes: Significance levers: *** - 0.001, ** - 0.01, * - 0.05

Table 3. Econometric estimation for Average Error

	1	**2**	**3**	**4**
Intercept	8.02 (5.39)	6.99 (4.26)	11.71* (5.2)	4.45** (1.45)
gender - M	-2.99 (4.83)	-4.96 (3.82)	-4.03 (3.86)	---
role defender	2.98 (4.27)	0.81 (3.38)	-0.05 (3.42)	---
role declarer	2.39 (5.23)	3.11 (4.13)	2.23 (4.15)	---
incentive	-2.49 (3.77)	1.11 (2.99)	---	---
number of overtricks	---	-17.04*** (1.6)	-16.92*** (1.62)	-16.82*** (1.58)
tournament 16052019	---	---	-7.27	---

			(4.62)	
tournament 17062019	---	---	-3.96 (4.16)	---
tournament 25042019	---	---	-5.9 (4.37)	---
R^2	0,01	0,38	0,39	0,38
R^2 adj.	-0,01	0,37	0,37	0,37
Number of observations	190	190	190	190

Notes: Significance levers: *** - 0.001, ** - 0.01, * - 0.05

Table 4. Econometric estimation for Absolute Error

	1	**2**	**3**	**4**
Intercept	30.85*** (1.98)	30.65*** (1.81)	32.4*** (1.95)	33.14*** (1.42)
gender - M	5.26** (1.72)	4.76** (1.57)	4.2** (1.58)	4.86** (1.57)
role defender	3.43* (1.63)	2.46 (1.48)	---	---
role declarer	1.91 (1.91)	1.73 (1.74)	---	---
incentive	1.64 (1.37)	2.31 (1.25)	---	---
number of overtricks	---	-8.6***	-8.58***	-8.6***

		(0.4)	(0.4)	(0.4)
tournament 16052019	---	---	4.85* (1.98)	---
tournament 17062019	---	---	-1.09 (1.84)	---
tournament 25042019	---	---	1.93 (1.87)	---
R^2	0,01	0,17	0,17	0,17
R^2 adj.	0,01	0,17	0,17	0,17
Number of observations	2318	2318	2318	2318

Notes: Significance levers: *** - 0.001, ** - 0.01, * - 0.05

Table 5. Econometric estimation for Average Absolute Error

	1	**2**	**3**	**4**
Intercept	37.4*** (2.97)	37.07*** (2.78)	35.4*** (3.39)	37.4*** (0.95)
gender - M	3.6 (2.67)	2.96 (2.49)	2.33 (2.52)	---
role defender	-2.46 (2.36)	-3.16 (2.21)	-2.56 (2.23)	---
role declarer	-4.46 (2.88)	-4.23 (2.7)	-3.61 (2.71)	---
incentive	-0.28	0.88	---	---

	(2.08)	(1.95)		
number of overtricks	---	-5.5*** (1.05)	-5.59*** (1.05)	-5.48*** (1.04)
tournament 16052019	---	---	5.12 (3.02)	---
tournament 17062019	---	---	1.2 (2.71)	---
tournament 25042019	---	---	2.4 (2.85)	---
R^2	0,02	0,15	0,17	0,13
R^2 adj.	0	0,13	0,13	0,12
Number of observations	190	190	190	190

Notes: Significance levers: *** - 0.001, ** - 0.01, * - 0.05

Discussion

The general conclusion from our exercise is that incentives do not reduce forecasts errors in bridge. That would suggest that poor calibration is due to inherent intellectual limitations, not due to insufficient motivation to think carefully. Then again, our results could also arise because the incentives were relatively. It also cannot be excluded that conscious effort plays some role, but players were relatively motivated to predict correctly even without the incentives.

The fact that the role played no role (defenders forecasting similarly to declarers) is interesting, given the voluminous literature on overconfidence and, more generally, positivity bias. One explanation is some sort of psychological hedging: a declarer may want to avoid a situation in which (s)he loses a contract (s)he deemed as very likely and similarly for the defenders.

References

Grieco, D., & Hogarth, R. M. (2009). Overconfidence in absolute and relative performance: The regression hypothesis and Bayesian updating. *Journal of Economic Psychology*, *30*(5), 756-771.

Hoelzl, E., & Rustichini, A. (2005). Overconfident: Do you put your money on it?. *The Economic Journal*, *115*(503), 305-318.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, *39*(1), 98-114.

Krawczyk, M. (2012). Incentives and timing in relative performance judgments: A field experiment. *Journal of economic psychology*, *33*(6), 1240-1246.

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology*, *18*, 455–463.

Yates, J. F., Lee, J. W., & Bush, J. G. (1997). General knowledge overconfidence: cross-national variations, response style, and “reality”. *Organizational behavior and human decision processes*, *70*(2), 87-94.

Appendix A: Stimuli

Instructions

Ladies and Gentlemen,

we represent the University of Warsaw and we would like to conduct a study during today's tournament, exclusively for scientific purposes. When the auction ends, players often wonder if the contract can be made. We would like to ask you to write down your prediction, expressed in percentage points. There are cards on the tables, please have a look but do not write anything yet.

For each deal, we will ask you to write down your pair number, indicate if your surname comes, alphabetically, before or after that of your partner (or your first name if both of you have the same surname). We will also ask you to write down the deal number and mark your position (N / E / S / W). This information may best be entered at your convenience, already during the auction.

And now the most important thing: **we will ask each of you (no matter if you are the declarer, the dummy, or one of the defenders) as soon as the auction ends, i.e. before the first lead, to enter your prediction: what is the chance that the declarer makes the contract.** This prediction should be based on your own cards and the course of the auction. For example, when the forecaster thinks the contract is certain, (s)he will enter 100%; (s)he will type in e.g. 60% or 30% if it can probably be made with favorable distributions (and the forecaster does not know if they are indeed favorable); she will write down 10% or less if it is very doubtful. It is not important whether there are overtricks or not (in case the contract is made) or if it's one down or two down etc. (in case it is not). We ask you, if possible, to enter this prediction in each hand, just before the first lead. If you are slightly late, please indicate in the appropriate field that predictions were made after the first lead. If you do not make it during the first trick, please leave the card empty.

Please fold the completed card in half and put it aside. The experimenter will collect the cards while the hand is being played.

Each of you will enter your own predictions: please do not reveal them and do not discuss them with other players. After all, hardly anyone wants to reveal it to his/her opponents how strong his/her cards are.

As a token of appreciation for your contribution to the study, after the tournament we will randomly select three participants and each of them will receive a prize of 100 PLN. [No Incentives Treatment]

As a token of appreciation for your contribution to the study, after the tournament we will randomly select three participants. Each of them will receive a monetary prize, ranging from 0 PLN to 200 PLN, depending on how accurate his/her predictions are. [Incentives Treatment]

If you wish, please also keep the pencil [a nice custom-made pencil with the symbols of the four bridge suits] as a souvenir. By contrast, we would like to use this information sheet the next time. Thank you very much for your cooperation!

Prediction card

Pair number ……..
My surname in the alphabet comes

○ before ○ after

that of my partner.
Deal number (please copy from the box) :

____

My position at the table:

○ N ○ E ○ S ○ W

I predict that the declarer will make the contract with probability __________%
[please do NOT fill it in after the first trick. Do so before the first lead if possible. If you have made your prediction already after the first lead (but before the end of the first trick) please mark here: □□

Was there a double on the final contract?

○ No ○ Yes, from player ______

Final survey

Pair number: □

In the alphabet your name comes: ○ before ○ after that of your partner

In comparison to your typical level of play, in today's tournament you played ...?

a) definitely worse than usual

b) rather worse than usual

c) more or less as usual

d) rather better than usual

e) definitely better than usual

f) hard to say

For each element of the game, please select the most accurate statement:

Compared to the average player of today's tournament, **I bid**…

a) much worse

b) a bit worse

c) more or less the same

d) a little better

e) much better

f) hard to say

Compared to the average player of today's tournament**, as a declarer I play**…

a) much worse

b) a bit worse

c) more or less the same

d) a little better

e) much better

f) hard to say

Compared to the average player of today's tournament, **I defend**…

a) much worse

b) a bit worse

c) more or less the same

d) a little better

e) much better

f) hard to say

For how many years, more or less, have you played in bridge tournaments?	
Approximately how many hours a week do you spend on bridge?	

a)

How would you describe your level of play on the scale below?

Complete novice	1	2	3	4	5	6	7	8	9	10	World class

Are you rather avoiding the risk or willing to take risks while playing bridge?

Risk avoidance	1	2	3	4	5	6	7	8	9	10	Full willingness to take risks

PLEASE TURN THE CARD

<u>The following questions do not apply to contract bridge.</u>

Are you generally a person who avoids risk or is ready to take risks?

Risk avoidance	1	2	3	4	5	6	7	8	9	10	Full willingness to take risks

What kind of driver are you?

a) Better than most drivers

b) Worse than most drivers

c) Hard to say

d) I do not drive a car at all

What kind of driver are you?

a) More cautious than most drivers

b) Less cautious than most drivers

c) Hard to say

d) I do not drive a car at all

<u>The following statements relate to investing money. For each of them, please indicate to what extent is true for you.</u>

When I invest, I choose specific assets myself (e.g. stocks of individual companies on the stock exchange).

a) it is completely untrue

b) it is rather untrue

c) hard to say / not applicable

d) it is rather true

e) it is completely true

When I invest, I eagerly choose more risky assets, as long as they can bring a higher profit.

a) it is completely untrue

b) it is rather untrue

c) hard to say / not applicable

d) it's rather true

e) it is completely true

When I invest, I stick to once selected assets for a long time. I rarely sell them to buy other ones.

a) it is completely untrue

b) it is rather untrue

c) hard to say / not applicable

d) it's rather true

e) it is completely true

Age: ____________ Years

Level and field of education: ____

Occupation:

Marital status:

a) single,

b) married,

c) widow/widower,

d) divorced,

e) in separation.

If you have any comments about today's survey, please enter them below or contact the experimenter.

____
_

____
_

Thank you for completing the survey!

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl