When Expert Advice Fails to Reduce the Productivity Gap: Experimental Evidence from Chess Players *

Elias Bouacida† Renaud Foucart‡ Maya Jalloul§

June 19, 2025

Abstract

We study the impact of external advice on the relative performance of chess players. We asked players in chess tournaments to evaluate positions in past games and allowed them to revise their evaluation after observing the answers of a higher or a lower-ability adviser. Although high-quality advice has the potential to serve as a "great equalizer," reducing the difference between higher- and lower-ability players, it did not happen in our experiment. One reason is that lower-ability players tend to pay a higher premium by sticking to their initial evaluation rather than following high-quality advice.

Keywords: decreasing differences, productivity gap, expert, advice, chess, control
JEL-Codes: C78, C91, C93, D91, J24, O33

*This project has received IRB approval from Lancaster University, the IRB file identification is FASSLUMS-2023-3790-LeXeL-3. We thank the Shamieh Chess Academy for their support in organizing the experiment, the Editor, Associate Editor, and two anonymous referees. we thank Alexander Matros and David Rietzke for taking the time to solve our positions. The paper also greatly benefited from the EEA mentorship programme and the comments of Danmo Lin and Paola Manzini, as well as from those of numerous colleagues, friends, and participants to seminars and workshop in Lancaster, Paris 8, JEI in Seville, the 1st Doctoriales ASFEE and 1st PBES who will recognize themselves.
The pre-registration is available at https://aspredicted.org/124_MSY and in Online Appendix F, the data is stored in https://doi.org/10.57745/UL502I, and the data analysis is available at https://decreasing-differences-ebouacida-92abd7c994ce472e25f3e62c7fbcaa.gitpages.huma-num.fr/
Previous versions of this paper circulated under the title *Decreasing Differences in Expert Evaluations: Evidence From a Field Experiment in Chess.*

†Université Paris 8, elias.bouacida@univ-paris8.fr
‡Lancaster University Management School, r.foucart@lancaster.ac.uk
§Lancaster University Management School, m.jalloul@lancaster.ac.uk

1 Introduction

Can offering high-quality advice help to narrow the productivity gap between higher and lower-ability workers? In theory, yes: lower-ability workers stand to gain more from relying on external advice. There is also plenty of empirical evidence showing that the benefits of collaborating with a high-productivity worker diminish as your own productivity increases. This has been observed across settings including garment factories (Hamilton, Nickerson, and Owan 2003; Adhvaryu et al. 2024) and university coursework (Fischer, Rilke, and Yurtoglu 2023). A study of the US labour market shows that lower-ability workers generally benefit more from teaming up with higher-ability partners (Herkenhoff et al. 2024). Chade and Eeckhout (2018) show theoretically how this logic applies to the context of matching expertise.

However, in a lab-in-the-field experiment with chess players participating in tournaments in Lebanon, we find no significant evidence that access to expert advice narrows the productivity gap. Our subjects reveal such a strong preference for sticking to their initial ideas and ignoring additional information that they miss out on a large share of the potential gains from advice. As a result, lower-ability players pay the highest premium for ignoring good advice.

In Summer 2023, we conducted an incentivized experiment in partnership with a local chess academy alongside chess tournaments across multiple locations in Lebanon. The main task required subjects to evaluate chess positions* and determine the corresponding *pawn advantage** – a measure of how much better a player's position is compared to the other.[1] For each position, subjects first gave their own evaluation by selecting one of four possible answers. We then provided them with an external adviser's evaluation and asked subjects to reassess the position.

One adviser is an International Master*, ranked among the top 6,000 players globally, with a higher rating than any of the subjects, who provided accurate advice for 75% of the positions. The second adviser is a casual player with no formal rating, placing him at the bottom of our subject pool, who provided accurate advice for only 15% of the positions. We define "higher-ability" subjects as those with an official chess rating in the top half of the sample and "lower-ability" subjects as those in the bottom half. Before receiving high-quality advice, higher-ability subjects had a correct answer rate of 41.2%, while lower-ability subjects had a rate of 32.9%.

After receiving high-quality advice, the correct answer rate increased to 50.8% (+9.6 percentage points) for higher-ability subjects, and to 42.5% (+9.6pp) for lower-ability

[1]For detailed explanations of chess terms and concepts used in the paper, see Appendix A. Terms identified with an asterisk are defined there.

subjects. In contrast, if higher-ability subjects had followed high-quality advice when it was beneficial on expectation, they could have increased their correct answer rate by an average of 22.0pp. Lower-ability subjects could have increased theirs by 30.0pp. Thus, the failure of expert advice to narrow the productivity gap in the experiment is largely due to the significantly higher premium paid by lower-ability subjects when ignoring high-quality advice.

The main novelty of our research is explicitly measuring the potential for narrowing the productivity gap in a context requiring domain expertise. Our subjects, ranging from less experienced chess players taking part in official tournaments to more experienced ones with official ratings, must reconcile their domain knowledge with potentially superior external advice. Unlike typical experimental subjects completing unfamiliar or routine tasks, our participants are receiving advice in an area where they already possess training and experience.

The decision to ignore one's own signal and follow the advice of others is typically studied in economics in the context of information cascades (Anderson and Holt 1997; Kübler and Weizsäcker 2004). This literature shows that subjects often prefer to rely on their own judgement, even when it is suboptimal, and tend to undervalue information discovered by others (Conlon et al. 2022; Weizsäcker 2010). In psychology, a large literature shows that individuals often give suboptimal weight to advice in their decision-making (Bailey et al. 2022; Bonaccio and Dalal 2006). This is linked to the preference for decision rights or control premium (Bartling, Fehr, and Herz 2014; Owens, Grossman, and Fackler 2014) and to the "illusion of control," (Langer 1975; Sloof and Siemens 2017) where individuals are overconfident when making their own decisions. We contribute to the literature on control and advice by providing results from a non-WEIRD (White, Educated, Industrialized, Rich, and Democratic) sample (Henrich, Heine, and Norenzayan 2010), as our subjects live in Lebanon, a Middle-Eastern country facing a banking and political crisis.

Recent research on advice given by Artificial Intelligence (AI) has also shown its potential to reduce the productivity gap in routine tasks for lawyers (Choi and Schwarcz 2025), programmers (Peng et al. 2023), writers (Noy and Zhang 2023), customer support (Brynjolfsson, Li, and Raymond 2025), and consultants (Dell'Acqua et al. 2023). This contrasts with studies on tasks where subjects perceive themselves as experts. For instance, Agarwal et al. (2023) found that radiologists often fail to optimally incorporate uncertain advice, and Otis et al. (2023) showed that among Kenyan entrepreneurs, advice benefits high performers but hinders low performers.

Our results align with work by Sandvik et al. (2020) and Sandvik et al. (2021), who found that in a workplace mentoring program, advice can decrease productivity differences between higher and lower ability workers. However, similar to our experiment, this potential was hindered by lower-ability workers' reluctance to opt into receiving advice.

In our experiment, advice involves providing subjects with experts' answers to the same questions that they are solving, with the expert's sole objective being to provide the correct answer. In contrast, strategic information transmission might lead to distorted advice if experts have different objectives than participants (Crawford and Sobel 1982) or if they aim to be perceived as competent (Pavesi, Scotti, and Argelli 2024; Renes and Visser 2024).

Finally, this paper follows a rich tradition of using chess players as a sample of highly qualified subjects to study human decision-making, including strategic behaviour in sequential games (Levitt, List, and Sadoff 2011), gender differences in risk-taking, behavioural preferences, and competitiveness (Backus et al. 2023; Engel 2025), the impact of time pressure on risk taking (Carow and Witzig 2024), social norms, role models and the gender gap (De Sousa and Niederle 2022; De Sousa and Hollard 2023; Dilmaghani 2021), the impact of work environment on productivity (Künn, Seel, and Zegners 2022; Künn, Palacios, and Pestel 2023), individual productivity over the life cycle (Strittmatter, Sunde, and Zegners 2020), and the role of superstars (Bilen and Matros 2023).

The rest of the paper is organised as follows. In Section 2, we describe our experimental protocol and procedures. We present the results in Section 3 and conclude in Section 4.

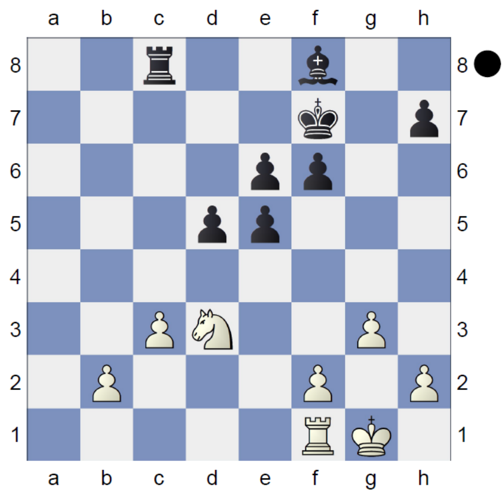
2 The experiment

We conducted the experiment during the Summer of 2023 in several cities in Lebanon, alongside tournaments organised by a local chess academy, on a sample of 102 subjects. Subjects were regular participants in tournaments. We describe their self-reported demographic characteristics in Table 3 in Appendix C.2. All subjects received the experimental material in both English and Arabic (the material is available in the Appendix G).

We recruited subjects prior to the tournament through the academy and covered their registration (approximately $5) as a participation fee. The experiment was conducted in a separate room while subjects were not engaged in competitive play. Each subject was randomly assigned to either the treatment with or without information on the adviser. There were two rounds of tasks, each corresponding to evaluating ten positions.

A position describes the arrangement of pieces at a given point in a game (Figure 1). Positions are evaluated using *pawn advantage*, a measure of which player (White or Black) is better positioned to win the game. We selected 20 positions from past chess games using the Chessbase Mega database 2023*, and selected half with a pawn advantage of

Figure 1: One position from our Experiment (Black's turn).



0.7 (a slight advantage) and half with 2.4 (a large advantage), either for Black (-2.4 and -0.7) or for White (0.7 and 2.4), loosely corresponding to a 60% and 80% probability of winning*. These pawn advantages are sufficiently distinct for skilled players to identify the correct one, while avoiding obviously winning positions.

The task was to identify the correct evaluation out of the four possible ones (in Figure 1, the correct answer is -0.7). Evaluating positions is a standard exercise familiar to any chess player sufficiently skilled to participate in an official tournament, and contemporary chess engines typically converge on similar pawn advantage evaluations, ensuring minimal ambiguity in determining the correct assessment.

In each round, subjects have 8 minutes to complete the first part of the answer sheet with ten evaluations. Then, they receive the evaluations from one adviser for the same positions. Subjects had 4 minutes to look back at their answers, compare with the advice, and complete the second part of the answer sheet with their possibly updated evaluations. We informed subjects that advisers evaluated positions in similar conditions.

In the *known adviser* treatment, we told subjects that the answers come from "a player with a rating of 2,335" (H-adviser) for one of the rounds of ten evaluations, and from "an unrated player who plays regularly for fun" (L-adviser) for the other one. In the unknown adviser condition, we told them in both rounds that "With equal probability, the player has a rating of 2,335, or it is an unrated player, who plays regularly for fun". The rating refers to the *Elo rating*,* the standard measure of chess performance and skill level.

After completing the two rounds of evaluations (ten pre-advice and ten post-advice evaluations per round, totaling 40 evaluations per subject), subjects filled out a brief demographic questionnaire along with questions regarding their stated preference for control from Burger and Cooper (1979). All sessions were administered by one of the co-authors of this study (Maya Jalloul), who read the experimental material aloud and made sure no one could cheat. The experimental material is available in the Online Appendix.

On top of the participation fee, one evaluation was randomly selected from the 40 answers and paid $10 if it was correct.[2]

In accordance with our pre-registration, we excluded one subject who did not provide answers after receiving the advice. Then, we divided the remaining sample of 102 subjects into two equal-sized groups based on their rating. As 54 subjects had a formal Elo rating, the results remain nearly identical when considering a dichotomy of rated/unrated subjects instead. We classified questions for which subjects did not provide answers as incorrect.[3] The Elo distribution of subjects is illustrated in Figure 5 in Appendix C.2. The average Elo rating of rated players is 1490, with the highest rated subject falling within the range 2,00-2,200. Therefore, it should be clear to all our subjects that the higher-ability adviser, with a rating of 2,335, is more likely to evaluate a position correctly than they are. Furthermore, as all subjects are tournament participants, it should be apparent that the lower-ability adviser (an unrated recreational player) does not possess higher ability than any subject.

3 Results

We show in Figure 2 the average share of correct answers in both ability groups (lower ability: l, higher ability: h), before observing advice, and after observing low-quality (L) and high-quality (H) advice. The figure pools both treatments.

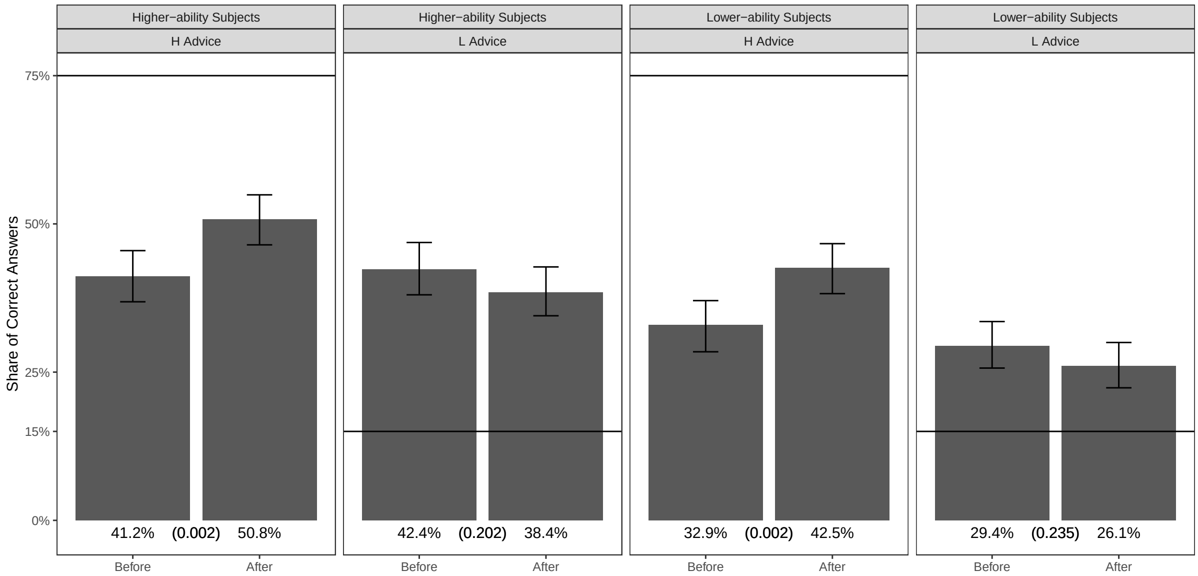
Before observing advice, higher-ability subjects evaluate 41.8% of the positions correctly, while lower-ability subjects evaluate 31.2% correctly (we show the results for each position in Appendix E.3). Our higher-quality adviser provided 75% of correct answers, and our lower-quality adviser only 15%, which is lower than the expected rate for random guessing (25%). As a result, the share of correct answers drops slightly, to 38.4% and 26.1% respectively after observing low-quality advice.

For positions with high-quality advice, the rate of correct answers increases from

[2]Given the difficult banking situation in Lebanon and the fact that some subjects were minors, we did not pay subjects directly in cash but with monetary vouchers for subsequent tournaments or other spending that day. We only knew the subject number, and not their identities. We communicated a list of payments and subject numbers to the organizing chess academy, who then processed the payments based on a list that they made allocating participant numbers to individuals.

[3]Not answering is a sub-optimal strategy, and it is a small percentage of all the answers, 2.0% before receiving the advice and 3.1% after. Using another hypothesis only marginally changes the proportions and does not change the results.

Figure 2: Performance across subject types before and after receiving advice.



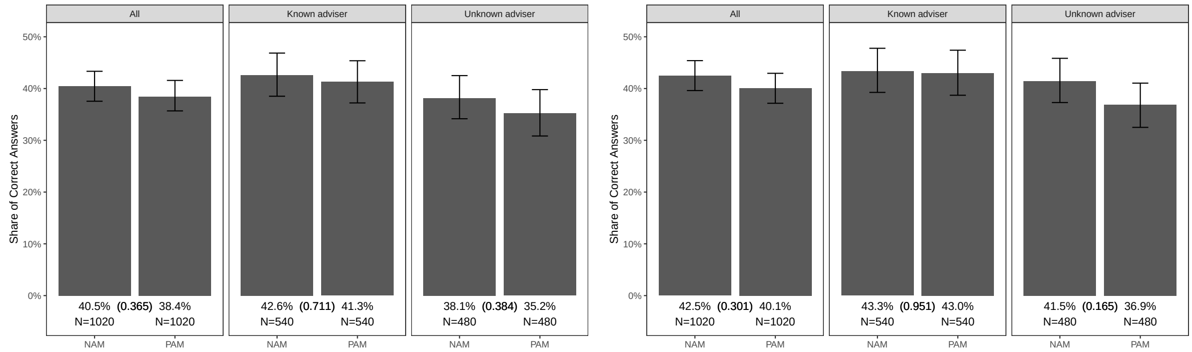
Note: The figure shows the share of correct answers by subject type (Higher-ability are the subject in the first half of the distribution of official ratings, Lower-ability are the others) before advice, and after receiving low-quality (L) or high-quality (H) advice. The data is pooled across treatments. Horizontal lines are the share of correct answers of the advisers. Each vertical bar represents the mean of 510 observations, and error bars show the 95% confidence intervals calculated using bootstrap resampling. In parentheses are p-values of the two-sided two sample equal proportion tests.

32.9% to 42.5% for lower-ability subjects, and from 41.2% to 50.8% for higher-ability subjects. The increase following high-quality advice is thus identical for both types of subjects, at 9.6 percentage points. However, if all subjects had followed all high-quality advice perfectly, our lower-ability subjects would have seen their share of correct answers after receiving high-quality advice increase by 8.3pp more than higher-ability subjects.

As a consequence, our pre-registered tests show no statistical evidence that expert advice reduces the productivity gap. In the main test, we compare the average share of correct answers when matching lower-ability subjects with low-quality advice and higher-ability subjects with high-quality advice (Positive Assortative Matching, PAM) to Negative Assortative Matching (NAM) of lower-ability subjects to high-quality advice and higher-ability subjects to low-quality advice. If expert advice reduces the productivity gap, lower-ability subjects should benefit more from high-quality advice than higher-ability subjects. Consequently, NAM should produce a higher share of correct answers than PAM (see Appendix B for a formalization).

We show in Figure 3a our main pre-registered proportion test of NAM versus PAM. None of the p-values reach statistical significance, by a wide margin.

Figure 3: Correct Answers with Negative (NAM) or Positive Assortative Matching (PAM).



(a) High vs. low-quality advice.

(b) High-quality advice vs. no advice.

Note: The figure shows the share of correct answers by matching type after receiving low-quality or no-advice (L) or high-quality (H) advice. NAM corresponds to the case where lower-ability subjects are matched to H-advice and higher-ability subjects to L-advice, and PAM is the opposite. Error bars show the 95% confidence intervals calculated using bootstrap resampling. In parentheses are p-values of the two-sided two-sample proportion test of equal proportion between NAM and PAM. N is the number of observation for each bar.

While we cannot rule out that expert advice reduced the productivity gap somewhat, our sample size was sufficient to detect any large effect. Following Vasilaky and Brock (2020), we look at the minimal detectable effect. With a sample size of 1,020 observations from both treatments, the proportion of correct answers with PAM was 38.4%. We could have detected a significant difference ($p < 0.05$) with a power of 0.8 if the NAM proportion had been 44.5%, compared to the observed 40.5%. As an additional robustness, we check how the rate vary and the non-significance of the results when excluding subjects who did not provide their demographic characteristics in Appendix D. Results remain unchanged.

In our alternative pre-registered test, we used the share of correct answers before receiving advice, rather than after receiving lower-quality advice as shown in Figure 3b. This test is particularly useful because our lower-quality advice is poor – 15% correct answers is worse than random guessing. We found no statistically significant evidence that expert advice reduced the productivity gap.

Overall, subjects tend to ignore advice: about three-quarters of choices remain unchanged after receiving advice (see Table 10 in Appendix E.1 for details by treatment and type). Unchanged answers can occur for several reasons, including when a subject's answer matches the adviser's, eliminating the need for modification. In Table 11 in Appendix E.1, we show that subjects kept their answers at least 93% of the time when they agreed with the adviser.

Table 1: How subjects react to the advice received when they disagree with it.

			Among those who disagree before				
Subjects	Treatment	Disagree	Keep	Follow	Closer	Further	N
Lower-ability	Know H	66.5	46.8	39.2	12.3	1.8	257
	Know L	72.2	78.8	14.0	7.3	0.0	248
	Unknown	70.6	64.3	27.3	6.3	2.2	452
Higher-ability	Know H	58.2	52.5	42.5	2.5	2.5	275
	Know L	73.2	83.9	9.5	4.0	2.5	272
	Unknown	67.0	79.7	15.3	3.3	1.7	449

Note: All numbers in the table are percentages. The first column, "Disagree," shows the share of subjects whose initial answer differs from the adviser's. Among those subjects, the second column is the share who Keep their initial answer. The third is those who Follow exactly the advice. The fourth those who change their answer to get Closer, but not identical, to the advice. The fifth those who change their answer even Further from the advice. N is the number of observations in each row. We remove from this table the missing answers because we have no distance from the answer for them. There are very few of them, as shown in Table 8 in Appendix C.3.2. We therefore slightly underestimate the disagreement percentage before receiving the advice.

Table 1 shows what happens when subjects' answers differ from those of the advisers. Lower-ability subjects ignore advice from unknown advisers 64.3% of the time, compared to 79.7% for higher-ability subjects. They also ignore (46.8%) or move further away (1.8%) from the higher-quality adviser about half of the time, slightly less than higher-ability subjects. In line with the findings of Alysandratos et al. 2020 on economic experts, we find no evidence that subjects can distinguish good from bad advice when the adviser's identity is unknown. In Appendix E.3, we show that the decision to follow advice is not significantly linked to the difficulty of solving the position either.

Participants mostly, and correctly, ignore low-quality advice. In line with Schultze, Mojzisch, and Schulz-Hardt 2017, some subjects felt compelled to incorporate even useless advice. Among higher-ability subjects, 9.5% updated their evaluations based on advice from advisers they should expect to be inferior, while the figure was 14.0% for lower-ability subjects.

We confirmed our main results in our regression analysis in Table 2. Following our pre-registration plan, we used both a binary classification of lower- and higher-ability subjects and a continuous measure based on the Elo rating. We conducted regressions for both known and unknown adviser treatments. As expected, the higher-ability adviser generally benefited subjects, and better-rated subjects were more likely to accurately evaluate positions. This effect was more pronounced in the Known Adviser treatment. Additionally, higher-rated subjects have a higher rate of correct answers, as expected. Individual characteristics mattered little, except for the relative ability of subjects.

The interaction term between advice quality and subjects' rating measures the reduction in the productivity gap due to access to expertise. As in the main tests using two subject categories, this effect was not significant.

Table 2: Regression analysis for the rate of correct answers.

	Known[1]	Unknown[1]	Known[2]	Unknown[2]
H adviser	0.396*	0.269*	0.315*	0.345*
	(0.160)	(0.133)	(0.157)	(0.135)
Elo[3]	0.306***	0.209**	0.326***	0.279***
	(0.065)	(0.067)	(0.066)	(0.076)
H adviser×Elo[3]	-0.180	-0.121	-0.112	-0.181+
	(0.121)	(0.103)	(0.116)	(0.096)
Female			0.088**	0.095*
			(0.032)	(0.048)
Age $\geq$ 30			-0.098+	-0.107+
			(0.068)	(0.064)
18 $\leq$ Age $\leq$ 29			-0.008	-0.037
			(0.033)	(0.047)
Control Index[4]			0.000	0.056
			(0.037)	(0.047)
Num. Obs.	1080	960	960	800
Num. Ind.	54	48	48	40
R-squared	0.145	0.080	0.171	0.092

Note: Fixed effects at position level with standard errors clustered at subject level. Dependent variable: rate of correct answers.

[1] Reference group: subjects receiving low-quality advice.

[2] Add controls for individual characteristics. Reference group: minor men (<18 years) receiving low-quality advice. Regressions on the same sample without controls reported in Appendix D.

[3] Elo rating scaled by 1,000 for coefficient readability..

[4] Control index measures stated preference for control.

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001.

To determine how much subjects sacrificed by ignoring advice, we compared their choices in our experiment with heuristics that either always follow or ignore certain types of advice. We did not pre-register the heuristics, nor the following analysis. Indeed, we did not foresee such low take-up of advice.

Our first heuristic, "Probability," represents the optimal choice subjects would make if they knew their likelihood of answering correctly (approximated by their share of correct answers) and the advisers'.[4] This heuristic represents the payoff-maximizing strategy subjects might adopt after numerous trials and errors.

For each subject, we applied this heuristic to determine how many correct answers they could have achieved by following it. Figure 4 quantifies the premium subjects paid for ignoring high-quality advice, defined as the difference in percentage points between the share of correct answers they could have achieved following our heuristic and their actual performance after receiving high-quality advice. Across treatments, lower-ability subjects could have increased their share of correct answers by 30.0pp after receiving high-quality advice by following the heuristic, compared to a 22.0pp increase for higher-ability subjects. The difference is statistically significant. Thus, lower-ability subjects pay a higher premium for ignoring high-quality advice, possibly due to overconfidence or preference for maintaining their original answer.

Although this method is imperfect and was not pre-registered, it illustrates the potential impact of advice as a great equalizer. In Appendix E.2, we present two additional heuristics with similar results: a "first best" heuristic where subjects select the correct advice when available for each question, and an "Elo" heuristic, where they follow only the advice of an objectively superior adviser.

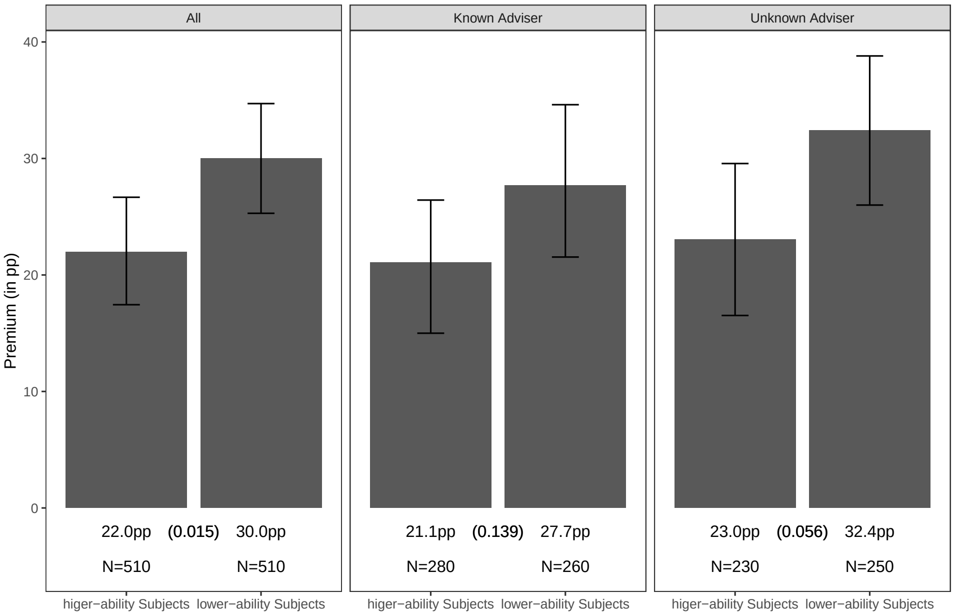
Finally, we constructed an index of stated preference for control by aggregating responses to questions adapted from Burger and Cooper (1979). We find no evidence that the stated preference for control correlates with the probability of subjects retaining their answers, after controlling for subject and position characteristics (see Table 12 in Appendix E.1).

4 Discussion and Conclusions

In this paper, we use a sample of subjects with specialist knowledge in a natural setting – chess players evaluating chess positions during a tournament – to learn more about the

[4] We approximate a subject i's probability of evaluating a position correctly, p_i, by their share of correct answers pre-advice. Similarly, we estimate the probabilities for experts, q_L and q_H, with $\bar{q} = \frac{q_L + q_H}{2}$ for unknown advice. This "probabilistic" way of incorporating advice follows a simple decision rule: if $p_i > q_H$, ignore all advice ; if $p_i \in (\bar{q}, q_H)$, only follow the known advice of H ; if $p_i \in (q_L, \bar{q})$, follow all advice except that of L ; and if $p_i < q_L$, follow all advice.

Figure 4: Premium paid for ignoring high-quality advice by ability and treatment.



Note: Premium paid by subjects receiving high-quality advice, defined as the difference between experimental payment and maximum possible payment from following our heuristic. Results shown pooled and by treatment group. Error bars: 95% confidence intervals (bootstrap method). Parentheses show p-values from two-sided t-tests comparing premiums between treatments. N is the number of observations for each bar.

"great equalizer" potential of advice. Although improving advice quality could benefit lower-ability subjects more, much of this potential is wasted because subjects stuck to their initial evaluations. This preference for relying on their own expertise harms lower-ability subjects the most, as they have the most to gain.

The general reluctance to take advice may stem from an intrinsic preference for autonomy and control. It may also result from anchoring and status quo bias, as subjects invest time in their answers, making them less likely to follow advice without careful consideration. We use an "independent-then-revise" procedure to measure how advice improves subjects' answers. Compared to offering advice before answering, it is unclear whether this approach leads to more or less advice-taking (Rader, Soll, and Larrick 2015). On one hand, answering first can lead to anchoring bias, where subjects prefer to keep their original answer. On the other hand, subjects who have the same answer as the adviser retain it, while seeing advice immediately can lead to a push-away effect, where subjects choose an answer just above or below the advice. Future experiments could study how confidence in their answers and further engagement with advice influences subjects' willingness to follow it. For instance, in politics, allowing communication between experts and subjects can increase the acceptability of advice (Zelizer 2022).

A limitation of our paper is that we do not identify the specific mechanism behind subjects' preference for autonomy and control, particularly among lower-ability players. One thing we do know is that chess provides an environment where subjects are aware of their relative abilities due to the rating system (González-Díaz, Palacios-Huerta, and Abuín 2024). Thus, we cannot attribute our results to lower-ability subjects not being aware of their abilities. A plausible explanation is that an intrinsic preference for control or overconfidence is negatively correlated with ability. This suggests that higher-ability subjects are also better at following advice. This correlation may not be exogenous: players who effectively incorporate advice during training may become the highest-rated in part because of their ability to take advice.

Our paper focuses on human advice. Comparing it with computer-based advice would have been challenging, as chess players have viewed algorithmic analysis as the gold standard since the landmark victory of chess engine Deep Blue over the then world champion Garry Kasparov in 1997. This is why we use chess engine evaluation of pawn advantages in our selected positions as the unambiguously correct answers. Further studies with specialists would benefit from comparing computer-based and human-based advice to determine which better reduces the productivity gap. Another limitation is our implicit assumption of a linear objective function for the social planner. We treat an improvement from 30% to 40% correct answers as equivalent to an improvement from 50% to 60%. If improving the outcomes of lower performers is prioritized, the conclusions of our experiment would be more optimistic.

References

Adhvaryu, Achyuta et al. (Oct. 2024). "No Line Left Behind: Assortative Matching Inside the Firm". In: *The Review of Economics and Statistics*, pp. 1–45. ISSN: 0034-6535. DOI: 10.1162/rest_a_01516.

Agarwal, Nikhil et al. (July 2023). *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology*. Working Paper 31422. National Bureau of Economic Research. DOI: 10.3386/w31422.

Alysandratos, Theodore et al. (2020). "The Expert and The Charlatan: an Experimental Study in Economic Advice". In: *Available at SSRN 3644653*. DOI: 10.2139/ssrn.3644653.

Anderson, Lisa R. and Charles A. Holt (1997). "Information Cascades in the Laboratory". In: *The American Economic Review* 87.5, pp. 847–862. ISSN: 00028282.

Backus, Peter et al. (2023). "Gender, competition, and performance: Evidence from chess players". In: *Quantitative Economics* 14.1, pp. 349–380. DOI: 10.3982/QE1404.

Bailey, Phoebe E et al. (2022). "A meta-analysis of the weight of advice in decision-making". In: *Current Psychology*, pp. 1–26. DOI: 10.1007/s12144-022-03573-2.

Bartling, Björn, Ernst Fehr, and Holger Herz (2014). "The intrinsic value of decision rights". In: *Econometrica* 82.6, pp. 2005–2039. DOI: 10.3982/ECTA11573.

Bilen, Eren and Alexander Matros (2023). "The Queen's Gambit: Explaining the superstar effect using evidence from chess". In: *Journal of Economic Behavior & Organization* 215, pp. 307–324. DOI: 10.1016/j.jebo.2023.09.002.

Bonaccio, Silvia and Reeshad S. Dalal (2006). "Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences". In: *Organizational Behavior and Human Decision Processes* 101.2, pp. 127–151. ISSN: 0749-5978. DOI: https://doi.org/10.1016/j.obhdp.2006.07.001.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond (Feb. 2025). "Generative AI at Work". In: *The Quarterly Journal of Economics* 140.2, pp. 889–942. ISSN: 0033-5533. DOI: 10.1093/qje/qjae044.

Burger, Jerry M and Harris M Cooper (1979). "The desirability of control". In: *Motivation and emotion* 3, pp. 381–393. DOI: 10.1007/BF00994052.

Carow, Johannes and Niklas M. Witzig (2024). "Time Pressure and Strategic Risk-Taking in Professional Chess". In: *Working Papers 2404, Gutenberg School of Management and Economics*.

Chade, Hector and Jan Eeckhout (2018). "Matching information". In: *Theoretical Economics* 13.1, pp. 377–414. DOI: 10.3982/TE1820.

Choi, Jonathan H. and Daniel Schwarcz (2025). "AI Assistance in Legal Analysis: An Empirical Study". In: *Journal of Legal Education* 73.2.

Conlon, John J et al. (2022). *Not learning from others*. Tech. rep. National Bureau of Economic Research. DOI: 10.3386/w30378.

Crawford, Vincent P and Joel Sobel (1982). "Strategic information transmission". In: *Econometrica: Journal of the Econometric Society*, pp. 1431–1451. DOI: 10.2307/1913390.

De Sousa, José and Guillaume Hollard (2023). "From Micro to Macro Gender Differences: Evidence from Field Tournaments". In: *Management Science* 69.6, pp. 3358–3399. DOI: 10.1287/mnsc.2022.4541.

De Sousa, José and Muriel Niederle (2022). *Trickle-down effects of affirmative action: A case study in France*. Tech. rep. National Bureau of Economic Research. DOI: 10.3386/w30367.

Dell'Acqua, Fabrizio et al. (2023). "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality". In: *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 24-013.

Dilmaghani, Maryam (2021). "The gender gap in competitive chess across countries: Commanding queens in command economies". In: *Journal of Comparative Economics* 49.2, pp. 425–441. DOI: 10.1016/j.jce.2020.09.006.

Engel, Julia (2025). *Chess, Gender, and Tournament Dynamics*. Tech. rep. Unpublished Working Paper.

Fischer, Mira, Rainer Michael Rilke, and B Burcin Yurtoglu (2023). "When, and why, do teams benefit from self-selection?" In: *Experimental Economics*, pp. 1–26. DOI: 10.1007/s10683-023-09800-2.

González-Díaz, Julio, Ignacio Palacios-Huerta, and José M. Abuín (Mar. 2024). "Personal Bests and Gender". In: *The Review of Economics and Statistics* 106.2, pp. 409–422. ISSN: 0034-6535. DOI: 10.1162/rest_a_01145.

Hamilton, Barton H, Jack A Nickerson, and Hideo Owan (2003). "Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation". In: *Journal of political Economy* 111.3, pp. 465–497. DOI: 10.1086/374182.

Henrich, Joseph, Steven J Heine, and Ara Norenzayan (2010). "Most people are not WEIRD". In: *Nature* 466.7302, pp. 29–29. DOI: 10.1038/466029a.

Herkenhoff, Kyle et al. (2024). "Production and Learning in Teams". In: *Econometrica* 92.2, pp. 467–504. DOI: https://doi.org/10.3982/ECTA16748.

Kübler, Dorothea and Georg Weizsäcker (2004). "Limited depth of reasoning and failure of cascade formation in the laboratory". In: *The Review of Economic Studies* 71.2, pp. 425–441. DOI: 10.1111/0034-6527.00290.

Künn, Steffen, Juan Palacios, and Nico Pestel (2023). "Indoor air quality and strategic decision making". In: *Management Science* 69.9, pp. 5354–5377. DOI: 10.1287/mnsc.2022.4643.

Künn, Steffen, Christian Seel, and Dainis Zegners (2022). "Cognitive performance in remote work: Evidence from professional chess". In: *The Economic Journal* 132.643, pp. 1218–1232. DOI: 10.1093/ej/ueab094.

Langer, Ellen J (1975). "The illusion of control". In: *Journal of personality and social psychology* 32.2, p. 311. DOI: 10.1037/0022-3514.32.2.311.

Levitt, Steven D, John A List, and Sally E Sadoff (2011). "Checkmate: Exploring backward induction among chess players". In: *American Economic Review* 101.2, pp. 975–990. DOI: 10.1257/aer.101.2.975.

Noy, Shakked and Whitney Zhang (2023). "Experimental evidence on the productivity effects of generative artificial intelligence". In: *Science* 381.6654, pp. 187–192. DOI: 10.1126/science.adh2586.

Otis, Nicholas et al. (2023). "The Uneven Impact of Generative AI on Entrepreneurial Performance". In: *Available at SSRN 4671369*. DOI: 10.2139/ssrn.4671369.

Owens, David, Zachary Grossman, and Ryan Fackler (2014). "The control premium: A preference for payoff autonomy". In: *American Economic Journal: Microeconomics* 6.4, pp. 138–161. DOI: 10.1257/mic.6.4.138.

Pavesi, Filippo, Massimo Scotti, and Nicola Argelli (2024). "Information and expertise". In: *The Elgar Companion to Information Economics*. Edward Elgar Publishing, pp. 202–223. DOI: 10.4337/9781802203967.00019.

Peng, Sida et al. (2023). *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*. DOI: 10.48550/arXiv.2302.06590.

Rader, Christina A, Jack B Soll, and Richard P Larrick (2015). "Pushing away from representative advice: Advice taking, anchoring, and adjustment". In: *Organizational Behavior and Human Decision Processes* 130, pp. 26–43. DOI: 10.1016/j.obhdp.2015.05.004.

Renes, Sander and Bauke Visser (2024). "Strategic Expert Committees and the Markets That Assess Them: A Laboratory Experiment". In: *Management Science*. DOI: 10.1287/mnsc.2021.03007.

Sandvik, Jason et al. (Apr. 2020). “Workplace Knowledge Flows”. In: *The Quarterly Journal of Economics* 135.3, pp. 1635–1680. ISSN: 0033-5533. DOI: 10.1093/qje/qjaa013.

— (2021). *Should workplace programs be voluntary or mandatory? evidence from a field experiment on mentorship*. Tech. rep. National Bureau of Economic Research. DOI: 10.3386/w29148.

Schultze, Thomas, Andreas Mojzisch, and Stefan Schulz-Hardt (2017). “On the inability to ignore useless advice”. In: *Experimental Psychology*. DOI: 10.1027/1618-3169/a000361.

Sloof, Randolph and Ferdinand A von Siemens (2017). “Illusion of Control and the Pursuit of Authority”. In: *Experimental Economics* 20.3, pp. 556–573. DOI: 10.1007/s10683-016-9499-7.

Strittmatter, Anthony, Uwe Sunde, and Dainis Zegners (2020). “Life cycle patterns of cognitive performance over the long run”. In: *Proceedings of the National Academy of Sciences* 117.44, pp. 27255–27261. DOI: 10.1073/pnas.2006653117.

Vasilaky, Kathryn N. and J. Michelle Brock (2020). “Power(ful) Guidelines for Experimental Economists”. In: *Journal of the Economic Science Association*. DOI: 10.1007/s40881-020-00090-5.

Weizsäcker, Georg (2010). “Do we follow others when we should? A simple test of rational expectations”. In: *American Economic Review* 100.5, pp. 2340–2360. DOI: 10.1257/aer.100.5.2340.

Zelizer, Adam (2022). “Talking shops: The effects of caucus discussion on policy coalitions”. In: *American Journal of Political Science* 66.4, pp. 902–917. DOI: https://doi.org/10.1111/ajps.12636.

Appendix

A Chess terminology

Below are chess terms and concepts used throughout the paper, in order of appearance.

Chess position A chess position is the arrangement of pieces on the chessboard at a specific moment during a game. In our experiment, subjects receive a diagram representing each position, with a colored circle indicating the side to move. Alternatively, a position could be described by providing the game's scoresheet, which includes annotated moves from the beginning of the game, or by using FEN notation, which captures all the details related to piece placement, the turn to move, castling rights, en passant rights, and move counters.

Elo rating The Elo Rating, created by Arpad Elo, computes the relative skill level of a player. It is perceived as a standard measure of a player's strength and ability. When two players play against each other in a tournament registered with the international chess federation FIDE, the winner gains Elo points, and the loser loses points. The number of points gained and lost depends on the rating difference and expected outcome. Any player with a rating strictly lower than 1,000 is considered unrated by the FIDE (and in our sample). As a rule of thumb, a difference of 100 points in the Elo rating means that the better-rated player is expected to win 5 out of 8 games. While Elo is an imperfect measure of ability, players take it seriously (see González-Díaz, Palacios-Huerta, and Abuín (2024)).

Pawn advantage Position evaluation is an important question for chess players and it estimates how favorable a chess position is for either side, whether preparing for a match or analyzing a game. Tournament players constantly evaluate positions using various approaches and factors. Pawn advantage quantifies this evaluation by measuring which side has the advantage in a given chess position and associating a numerical value to it.

Chess title The International Chess Federation (FIDE), along with some national organizations, awards chess titles as official recognition of a player's skill level and achievements. The most prestigious titles, in order of importance, are Grandmaster (GM), International Master (IM), FIDE Master (FM), and Candidate Master (CM), as well as their Women's counterparts: WGM, WIM, WFM and WCM.

ChessBase Mega database ChessBase is a chess software used for analyzing and studying games and managing databases. It permits storing and analyzing games and positions using chess engines. The ChessBase Mega database is a collection of chess games from historical matches through recent tournaments, updated annually.

Probability of winning Chess players typically avoid translating pawn advantages into winning probabilities, primarily because their focus during a game is evaluating positions and finding the best moves to secure an advantage. Additionally, chess has three possible outcomes (win, loss, or draw), making it more complex than a simple win/loss scenario. According to one measure (developed by Sune Fischer and Radu Pannan using 405,460 past games), a pawn advantage of 0.7 corresponds to a 60% probability of winning and of 2.4 to a 80% probability of winning – counting a draw as half a win. Our selected positions follow this statistical pattern: games with a ±2.4 pawn advantage resulted in 7 wins for the advantaged player, 2 draws, and 1 loss ; games with a ±0.7 pawn advantage resulted in 3 wins, 6 draws, and 1 loss.

Chess engine A chess engine is a computer program designed to analyze and play chess at a sophisticated level. It uses advanced search algorithms to find and rank moves, evaluate positions and calculate different variations.

Position selection Positions for the experiment come from Mega Database 2023, containing over 9.75 million games. A random number generator was used to choose the games, and we examined positions from the middlegame or early endgame within one of the following evaluation ranges: + or - 0.7, + or - 2.4. We aimed to include positions with either slight or significant advantage for White or Black. For the selected positions, we considered an engine analysis search depth of 25 to 30 plies (a ply is a single move made by one player), as this represents a reasonably strong engine evaluation.

B A simple theoretical framework

Consider two subjects l and h, with perfect information about their own probability of successfully solving a task p_i, $i \in \{l, h\}$, as well as the probability of the lower L and higher H ability advisers to do so $q_L < q_H$. When subjects do not know the identity of the adviser – but know both are equally likely – we denote by $\bar{q} = \frac{q_L + q_H}{2}$ this probability.

Unless all subjects follow (or ignore) all types of advice, we should observe strictly *decreasing differences* (what we call for simplicity in the main text "reducing the productivity gap") if subjects correctly infer the probabilities and maximize their expected probability of finding the correct answer.

Define by $f(i, j)$ the probability that subject i solves a task correctly after observing advice j and assume that $q_L < p_l < p_h < q_H$. If subjects want to maximize their probability of success and know the identity of the adviser, $f(l, L) = p_l$, $f(h, L) = p_h$, and $f(l, H) = f(h, H) = q_h$.

It is easy to see that in that case, the function displays decreasing differences:

$$f(l, H) - f(l, L) > f(h, H) - f(h, L),$$

as the expression simplifies to $p_l < p_h$. This statement is equivalent to saying that Negative Assortative Matching (NAM) of subjects to advisers yields a higher expected share of correct answers than Positive Assortative Matching (PAM):

$$\frac{f(l, H) + f(h, L)}{2} > \frac{f(l, L) + f(h, H)}{2}.$$

The same result holds when considering the case of unknown advisers if $p_l < \bar{q} < p_h$, so that type l subjects follow all advice and type h do not follow any. In that case, $f(l, L) = q_L$, $f(h, L) = f(h, H) = p_h$, and $f(l, H) = q_H$. The condition for decreasing differences is then $q_H > q_L$, and the difference between NAM and PAM is higher than with known advisers. The reason is that a good adviser then not only helps more the lower-ability subjects, but it also protects them from following bad advice. Finally, if $\bar{q} \geq p_h$ or $\bar{q} \leq p_l$, the differences are constant and the probability of a correct answer in NAM is the same as in PAM. This result is trivial, as it simply states that if all subjects follow all advice, they also solve all problems with the same probability, and if they ignore all advice, the quality of advice has no influence on their success.

By the same logic, we can compare advice from H and no advice at all, where $f(i, 0) = p_i$ is the probability of the answer of subject i being correct before advice. With known adviser, the result is identical to the one above, as $f(i, L) = p_i$ for both types of subjects. With unknown adviser, there are always decreasing differences unless all advice is ignored. If $\bar{q} < p_h$, the condition becomes $q_H > p_l$. If $\bar{q} \geq p_h$, it is $p_h > p_l$.

There are however two main biases and preferences that could influence our theoretical result of decreasing differences in the experiment. The first is that our subjects do not have full information on their probability of success and the one of their advisers. If lower-ability subjects are also more overconfident than higher-ability ones, they may benefit relatively less from advice. The second is preference for following their initial idea: if lower-ability subjects value more strongly keeping their first answer than higher-ability ones, they are less likely to follow advice for a given expected gain, decreasing the potential for advice to act as a great equalizer.

C Experiment Description

C.1 Recruitment

The exact dates of the tournaments are August 15, August 20, September 2, and September 17, 2023. Following the pre-registration, we stopped recruiting participants when we reached 100 subjects, so that we recruited a total of 103 subjects. Our total sample is however $n = 102$ as, following the pre-registration, we removed observations for which no choice were made and one of our subjects did not write anything in the second part of the answer sheet. No subject was caught cheating and therefore none was excluded for that reason.

C.2 Description of the Sample

Table 3 indicates that the majority of our subjects are young men. Figure 5 illustrates that while most of our subjects are rated, the mode category is being unrated. The number of unrated players means that the lower-ability group is mostly composed of unrated players.

Table 3: Self-declared demographic characteristics.

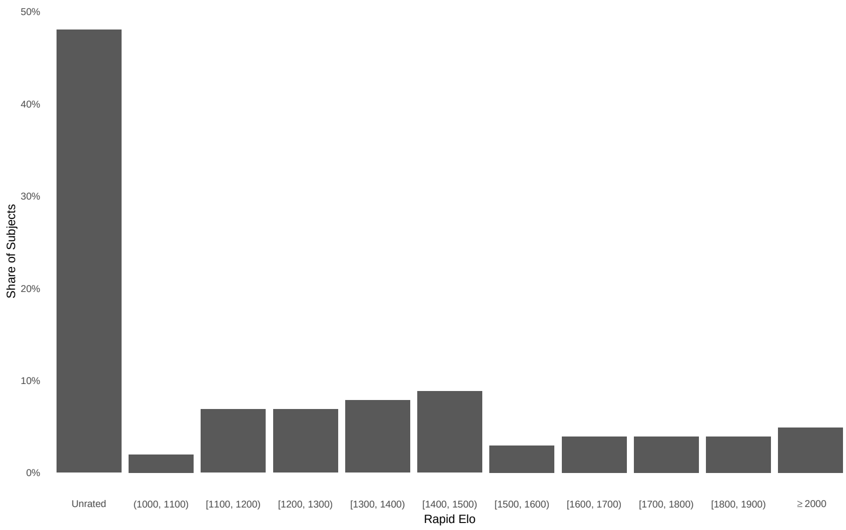
		Age				
		<18	18-29	≥ 30	Undeclared	Total
Gender	Female	6	3	1	-	10
	Male	27	38	13	-	78
	Undeclared	1	1	-	12	14
	Total	34	42	14	12	102

C.3 Randomization Checks

C.3.1 By Demographic Characteristics

In this section, we check that the sample is well randomized between the known and unknown adviser treatments. We first start with demographic characteristics. We find no significant difference in the proportion of male in the two treatments in Table 4. We look at the composition by age in the two treatments in Table 5. There seems to be relatively more young adults (age between 18 and 29) in the known adviser treatment. The p-value of the Fisher test of the proportion of young adults compared to the rest of

Figure 5: Distribution of Elo ratings among our subjects. The higher-ability adviser is rated above the upper limit.



the sample is however 0.07, which is not significant at conventional levels. Finally, Table 6 shows that there is no significant difference between lower and higher-ability subjects between treatments. If instead of using our higher/lower-ability distinction, we use the distinction rated/unrated, the picture is very similar.

Table 4: Gender composition by treatment.

Gender	Known Adviser	Unknown Adviser
Female	5	5
Male[1]	43	35
Undeclared	6	8

[1] The p-value of the Fisher test of equal proportion of male compared to the non-male in the two treatments is 0.487.

Table 5: Age composition by treatment.

Gender	Known Adviser	Unknown Adviser
<18	16	18
18-29	27	15
≥30	7	7
Undeclared	4	8

Finally, we show in Table 7 the composition of the treatments by sessions. There is

Table 6: Composition of the treatments by ability.

Ability	Known Adviser	Unknown Adviser
Higher	28	23
Lower	26	25

Note: The p-value of the Fisher test of equal proportion of higher ability participants in the two treatments is 0.843.

quite a large variation between the sessions in the number of subjects, but within sessions, the number of participants in each session is approximately the same.

Table 7: Composition of the treatments by dates.

Date	Known Adviser	Unknown Adviser
08/15/2023	12	10
08/20/2023	29	29
09/02/2023	10	7
09/17/2023	3	2

C.3.2 By Answer Status

For a final randomization check between treatments, we show the proportion of non answers before receiving advice by treatment in Table 8. The difference is significant, but it is because it is driven by two subjects in the unknown adviser treatment not answering to at least 10 questions. We exclude in the main statistics positions were we did not have answers before and after, so that these subjects did not influence our results.

Table 8: Number of non-answers by treatment.

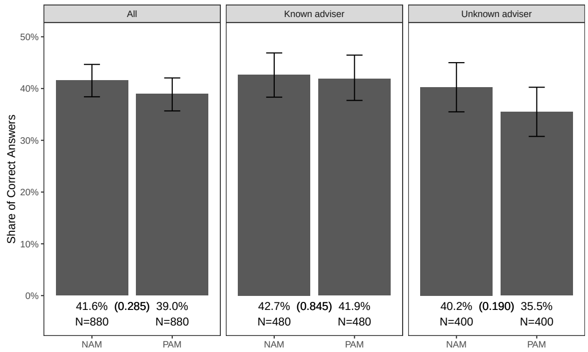
Adviser	Answered	Non-answered
Unknown	932	28
Known	1067	13

Note: The p-value of the Fisher test of equal proportion of answers is 0.007.

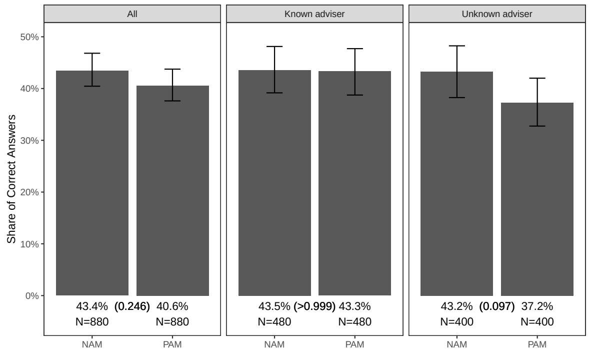
D Robustness Tests

Some subjects did not answer the demographic questionnaire; as per our IRB registration, providing this information was voluntary. As it may raise questions about the reliability of their answers, we proceeded to run our main test on a restricted sample where subjects answered all demographic questions. The results are shown in Figure 6. They are very similar to the results given in Figure 3. The main difference is that, without advice, the difference between NAM and PAM marginally reaches the 10% significance threshold.

Figure 6: Negative and Positive Assortative Matching (NAM and PAM) share of correct answers after receiving advice, pooled and by treatment, restricted to subjects who answered all the demographic questions.



(a) With the low quality advice.



(b) When replacing low quality advice with no advice.

Note: Share of correct answers by matching type after receiving low-quality or no advice (L) or high-quality (H) advice. Error bars: 95% confidence intervals (bootstrap method). In parentheses are p-values from two-sided two-sample proportion tests comparing NAM and PAM within each advice condition. N is the number of observations for each bar.

We also ran the same regressions as in Table 2 using this restricted sample. The results, shown in Table 9, yield similar conclusions: while the coefficient magnitudes differ slightly, all signs and significance levels remain unchanged.

Table 9: Regression analysis for the rate of correct answers on the restricted sample.

	Known	Unknown
H adviser	0.315*	0.345*
	(0.157)	(0.135)
Elo[1]	0.283***	0.201**
	(0.068)	(0.072)
H adviser×Elo[1]	-0.112	-0.181+
	(0.115)	(0.096)
Num. Obs.	960	800
Num. Ind.	48	40
R-squared	0.164	0.081

Note: Fixed effects at the position level with standard errors clustered at the subject level. Dependent variable: Rate of correct answers. Reference group: subjects receiving low-quality advice.

[1] Elo rating scaled by 1,000 for coefficient readability.

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001.

E Additional Results

E.1 Keeping Answers

Table 10 shows that subjects correctly change their answers more after seeing high-quality advice while knowing it. They fail to change them as often as they should, however. As expected, lower-ability subjects change answers more often than higher-ability subjects, partly because their initial responses are more likely to differ from high-quality advice recommendations. When subjects agree with their adviser's recommendation, they typically retain their original answers (Table 11). The regression results in Table 12 confirm that higher-Elo subjects are more likely to stick with their initial responses than lower-rated players. The "Know adviser" dummy captures situations where subjects know they received low-quality advice – in these cases, they correctly retain their answers more frequently. Conversely, the interaction terms show that subjects changes significantly more their answers when they know the advice is high-quality. Regression (2) indicates that individual characteristics beyond subject ability do not significantly influence these patterns.

Table 10: Share of identical answers for lower- and higher-ability subjects after observing different qualities of advice.

Subject	Low-quality advice	High-quality advice	Unknown advice
Lower-ability	80.4%	61.5%	69.4%
Higher-ability	85.0%	69.6%	83.5%

Table 11: Percentage of kept answers when subjects agree with the adviser.

			When Agree Before[2]		
Subjects	Treatment	Agree[1]	Keep	Change	N[3]
Lower-ability	Know H	33.5	93.0	7.0	257
	Know L	27.8	98.6	1.4	248
	Unknown	29.4	94.0	6.0	452
Higher-ability	Know H	41.8	96.5	3.5	275
	Know L	26.8	95.9	4.1	272
	Unknown	33.0	97.3	2.7	449

[1] Percentage of identical pre-advice answers with the adviser. To align with Table 1, we remove for this computation the missing answers because we have no distance from the answer for them. We therefore slightly overestimate the agreement percentage before receiving the advice.

[2] Percentage of kept or changed answer conditional on pre-advice answer being identical to the adviser's.

[3] Number of answers in this row.

Table 12: Regression for keeping the answer after receiving the advice.

	Share answer kept	
	(1)	(2)
Distance Correct[1]	-0.041***	-0.045***
	(0.008)	(0.008)
H adviser	-0.019	-0.008
	(0.034)	(0.037)
Known adviser	0.078+	0.047
	(0.042)	(0.040)
Elo[2]	0.177***	0.170***
	(0.052)	(0.045)
H adviser×Known adviser	-0.168***	-0.169**
	(0.048)	(0.052)
Control Index		0.036
		(0.036)
Female		-0.046
		(0.034)
Age $\geq$ 30		0.044
		(0.051)
18 $\leq$ Age $\leq$ 29		0.007
		(0.039)
Num. Obs.	1999	1749
R-squared	0.072	0.080

Notes: Fixed effects at the position level with standard errors clustered at the subject level. Dependent variable: share of kept answers. Reference group: subjects receiving low-quality advice.

(2) adds demographic controls but restrict the sample, reference group becomes minor men (<18 years old) receiving low-quality advice.

[1] Absolute distance from the correct answer in pawn advantage.

[2] Elo rating scaled by 1,000 for coefficient readability.

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

E.2 Alternative Heuristics

In Section 3, we presented a counterfactual "probabilistic" heuristic of what subjects could have achieved if they were aware of their average probability of being correct. Here, we propose two alternative heuristics.

First, we start with a highly unrealistic "first-best" heuristic. For each evaluation, subjects can switch their choice to align with the adviser's recommendation, assuming that they know the correct position after receiving the advice. Essentially, this assumes that subjects recognize the correct answer once they see the advice but are limited to choosing between their initial decision and the adviser's suggestion.

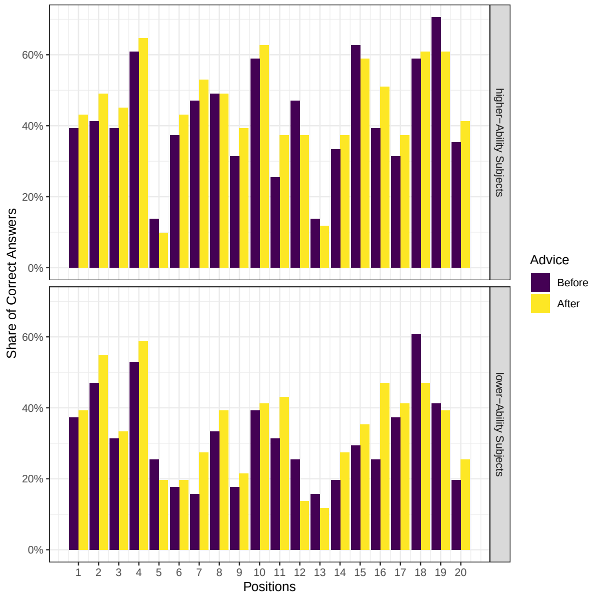
Our second heuristic, the "Elo" heuristic, is a rule-of-thumb that involves accepting advice only from someone objectively better. For higher-ability subjects, this means following only known higher-quality advice, while for lower-ability subjects, it means ignoring only known lower-quality advice. This approach is straightforward and relies on information available to the subjects beforehand. However, it is simplistic and disadvantages higher-ability subjects who might benefit from following some unknown advice.

Table 13 shows that regardless of the heuristic applied, lower-ability subjects always pay a higher premium than higher-ability subjects. The difference is not always statistically significant, particularly with the first-best heuristic which inherently equalizes more than the other two heuristics.

E.3 Difficulty of Positions

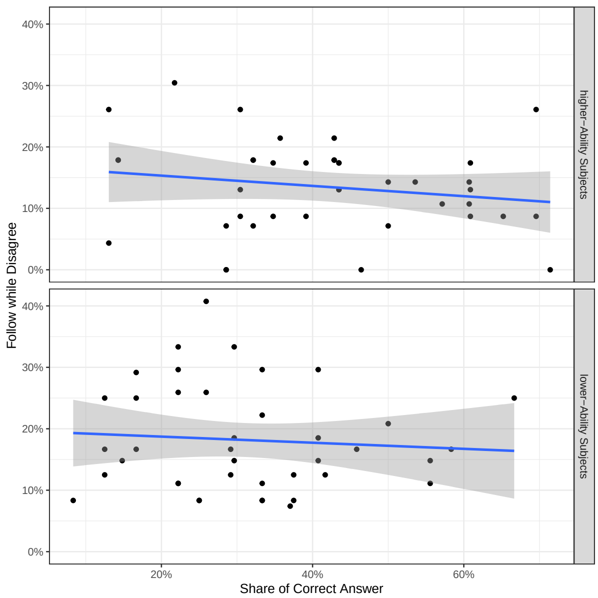
Figure 7 shows that not all positions lead to the same share of correct answers. In general, higher- and lower-ability subjects show the same pattern: there does not seem to be any differential effect of the position on the share of correct answers. One explanation for this correlated pattern could be that some positions are easier to evaluate and some are harder. In itself, it is not a problem for our analysis as we control for the positions by using positions fixed effects. It could be a problem, however, if there is a differential effect on following the advice given. Figure 8 shows that there is no such effect. If there was one, the slope of the linear approximation would differ between higher and lower-ability subjects, which is not the case, as the regression in Table 14 shows. If it were the case, then the interaction terms between the share of correct answers before and being a higher-ability subject would be significant. To conclude, while there is a variation in the share of correct answers by position and, presumably, in the perceived difficulty of each position for subjects, it does not influence how they followed advice and therefore our results.

Figure 7: Share of correct answers before and after receiving the advice, by position and ability.



Note: There are 51 observations for each average computation.

Figure 8: Share of subjects following advice given while disagreeing before with it by the share of correct answer before.



Note: The line represents the linear approximation with 95% confidence interval. Each dot is the average for a given position.

Table 13: Difference (in percentage points) between the average share of correct answers of lower- and higher-ability subjects having received high-quality advice, following our heuristics and in the experiment.

		Subjects			
Heuristic	Treatment	Lower-ability	Higher-ability	P-value[1]	N[2]
Probabilistic	Unknown	32.4	23.0	0.056	480
	Known	27.7	21.1	0.139	540
	All	30.0	22.0	0.015	1080
Elo	Unknown	36.8	-5.7	< 0.001	480
	Known	27.7	21.1	0.139	540
	All	32.2	9.0	< 0.001	1020
First-best	Unknown	43.2	37.8	0.240	480
	Known	34.6	28.9	0.168	540
	All	38.8	32.9	0.055	1020

Note: Premia paid for ignoring high-quality advice are given in percentage points.

[1] P-value of the two-sided two sample t-test of equal premium between the lower- and higher-ability subjects.

[2] N is the number of observations in each sample.

Table 14: The slope of the share of correct answers per position are not significantly different between higher and lower-ability subjects.

	(1)
(Intercept)	0.197
	(0.032)
Share Correct Before	-0.050
	(0.094)
higher-ability subjects	-0.027
	(0.048)
Share Correct Before×h-Ability Subjects	-0.034
	(0.123)
Num. Obs.	80
R-squared	0.096

OLS regression for the share of subjects following advice after disagreeing with it.
+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001,

F Pre-Registration (for Online Publication)

The pre-registration is time-stamped and available on AsPredicted (https://aspredicted.org/124_MSY).

Below, **bold text** indicates standard pre-registration questions, plain text shows our registered responses, and *italicized text* references where the corresponding analysis appears in this paper. For consistency, we have edited the notation in the pre-registration below to match the paper's terminology.

- **Have any data been collected for this study already?**

  No, no data have been collected for this study yet.

- **What's the main question being asked or hypothesis being tested in this study?**

  Using an incentivized experiment on a sample of chess players of various abilities, we want to learn how experts learn from each other. Our main question is whether the input of experts displays decreasing differences, in the sense that the input of high-quality experts benefits more to lower quality ones than to high quality ones. We also want to look at whether information about the status of the expert affects this result.

- **Describe the key dependent variable(s) specifying how they will be measured.**

  We will classify our subjects into two categories. Category l will be the 50% lower rated subjects, mostly those who do not have an Elo rating. Category h will be the higher rated subjects. For robustness, we will also try different separations of types using a cutoff rated/unrated. If more than 50% of the subjects are unrated, we will use the cutoff rated/unrated as the main distinction.

  The main outcome of the experiment is the accuracy of the answers of our subjects. We ask them to evaluate positions in chess, by picking one of four possible "pawn advantages" (a measure of how well positioned a player is at a certain point of the game). They then receive the evaluation of one of our experts for the same position, and they are free to update their choice.

  For each of the observations (predictions), we will measure it as a binary correct/incorrect answer. We will have the answers of the subjects before and after receiving the advice of an expert.

  The advice of the experts has been provided by two chess players that we know. One is an amateur, unrated, that we expect to be close to the worse players of the tournament. One is highly rated, with an Elo that we expect to be close to the top players of our tournament. They have evaluated the positions in the same conditions as the actual experiment.

We provide these descriptive statistics in Figure 2.

- **How many and which conditions will participants be assigned to?**

  We will allocate randomly our subjects to one of two treatments. In the first one, subjects will have information about the quality of the expert advice they receive (their Elo rating). In the second one, they will not have this information. Then, each subject will play a round of 10 predictions, first without input and then using the predictions from one expert, and one round of 10 predictions first without input, then using the predictions from the other expert. We will randomize the order of the expert advice received.

- **Specify exactly which analyses you will conduct to examine the main question/hypothesis.**

  Decreasing differences means that the quality of evaluations following the advice of a good expert improve more when the player is of low type. Denote by l, h the ability of the players. By L, H the ability of the expert. And by 0 the case without expert advice. Denote by f the quality of an evaluation. Decreasing differences implies that: $f(l,H) - f(l,L) > f(h,H) - f(h,L)$ We want to measure this separately for the treatment with and without information. Rearranging the above equation, we compare $f(l,L) + f(h,H)$ and $f(l,H) + f(h,L)$.

  We provide these results on Figure 3.

- **Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.**

  We will exclude observations where no choice or more than one choice were made among the four possible pawn advantage. We will exclude subjects who tried to talk to each other, use their phone or cheat by any other way.

  We excluded one subjet who did not provide answer after receiving the advice. No one was caught cheating.

- **How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.**

  We hope to have at least 100 participants, not much more, and around 50 unrated and 50 rated. This makes it 100*40=4,000 observations, with 2,000 being used for our main analysis (after observing an expert's advice). We co-organise a chess tournament in Lebanon on August 20, where we should have most our participants. We will also recruit subjects in other chess tournaments by the same organizer.

- **Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)**

  We want to look at:

  (i) The impact of the low ability experts, comparing $f(l, L) + f(h, L) and f(l, 0) + f(h, 0)$.

  We provide this result on Figure 3b

  (ii) How much does the median prediction for each of the 20 games improved after adding the expert advice with and without disclosing status.

  We provide this result in Appendix E.3.

  (iii) Preference for control, Regression analysis with a continuous measure of the rating, and a measure of the distance to the correct answer.

  We provide a regression analysis with the different controls in Table 2. We look at the role of distance to the correct answer on Table 12.

G Instructions (for Online Publication)

We include in the following pages the instructions given to subjects. In each session of the experiment (i.e., each tournament), subjects could be in one treatment or the other. Subjects were not allowed to talk to each other during the experiment and we made sure of that.

FOR ONLINE PUBLICATION ONLY: EXPERIMENTAL INSTRUCTIONS IN THE KNOWN ADVISOR TREATMENT

Response Sheet - ورقة الاجابة

Participant number – رقم المشترك(ة):A1___________

ELO - التصنيف:

Your predictions - التوقعات

Round 1 – الجولة الاولى:

رقم الوضع Position Number	Part 1 - الجزء الأول				Part 2 - الجزء الثاني			
	-2.4	**-0.7**	**+0.7**	**+2.4**	**-2.4**	**-0.7**	**+0.7**	**+2.4**
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								

Round 2 – الجولة الثانية:

رقم الوضع Position Number	Part 1 - الجزء الأول				Part 2 - الجزء الثاني			
	-2.4	**-0.7**	**+0.7**	**+2.4**	**-2.4**	**-0.7**	**+0.7**	**+2.4**
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								

(Please complete both sides of the sheet)

الرجاء تعبئة جانبي الورقة

Additional Info – معلومات إضافية

Age - العمر:

Gender - الجنس:

How much do you agree with the following statements - ما مدى موافقتك على العبارات التالية ?

	Strongly disagree أعارض بشدة	Disagree أعارض	Neither agree nor disagree لا أوافق ولا أعارض	Agree أوافق	Strongly agree أوافق بشدة
I try to avoid situations where someone else tells me what to do. أحاول تجنب المواقف التي يقول لي فيها شخص آخر بما يجب القيام به.					
I prefer to be a leader rather than a follower. أفضل أن أكون قائدًا وليس تابعًا.					
I enjoy making my own decisions. أنا أستمتع باتخاذ قراراتي بنفسي.					
I would rather someone else took over the leadership role when I'm involved in a group project. أفضل أن يتولى شخص آخر الدور القيادي عندما أشارك في مشروع جماعي.					
There are many situations in which I would prefer only one choice rather than having to make a decision. هناك العديد من المواقف التي أفضل فيها خيارًا واحدًا فقط بدلاً من الاضطرار إلى اتخاذ قرار.					

A1i

Here are ten positions that occurred in real chess games which have been chosen from a dataset of previous games from the Mega Database 2023.

We will ask you to evaluate 20 games over two rounds: 1 and 2. We will pick one of your evaluations at random and you will receive a voucher of $10 if your answer was correct.

Please complete **Round 1, Part 1** of the Response Sheet by indicating for each game your best estimate of the pawn advantage, which can be +0.7, -0.7, +2.4, or -2.4. Please check the box corresponding to your choice (only one possible answer). Note that the positions have a pawn advantage of ±0.7 and one of ±2.4 with equal probability.

Once you have completed **Round 1, Part 1**, please wait for the experimenter to give you the next set of instructions.

You have a total of 8 minutes to complete this part.

فيما يلي عشرة أوضاع حصلت في جولات شطرنج حقيقية وقد تم اختيارها من مجموعة بيانات للألعاب السابقة من Mega Database 2023.

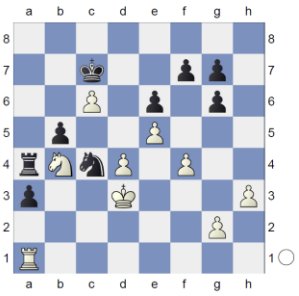
سنطلب منك تقييم 20 وضع على مرحلتين: الجولة الأولى والجولة الثانية. سوف نختار أحد تقييماتك بشكل عشوائي وستتلقى قسيمة بقيمة 10 دولارات إذا كانت إجابتك صحيحة.

يرجى تعبئة **الجولة 1، الجزء 1** من ورقة الإجابة بالإشارة إلى أفضل تقدير لديك لكل وضع حسب أفضلية ال pawn advantage والتي يمكن أن تكون 0.7+ أو 0.7- أو 2.4+ أو 2.4-. يرجى تحديد المربع المقابل لاختيارك (إجابة واحدة فقط ممكنة). ملاحظة: من المحتمل أن يكون الوضع مع أفضلية 0.7±، **أو** 2.4±، مع احتمالية متساوية.

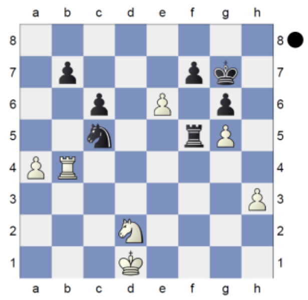
بمجرد الانتهاء من الجولة 1، الجزء 1، من فضلك انتظر أن يعطيك المشرف المجموعة التالية من التعليمات.

لديك 8 دقائق لإكمال هذا الجزء.

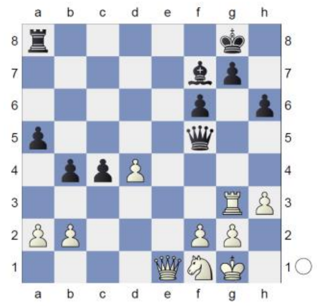
1



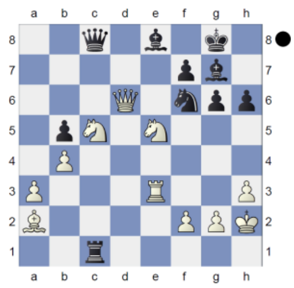
2



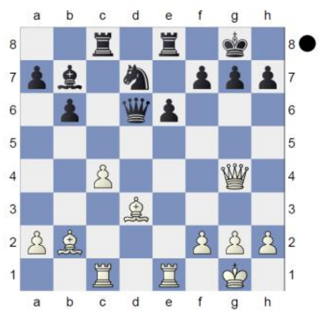
3



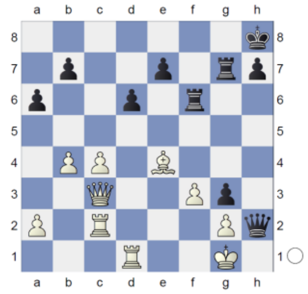
4



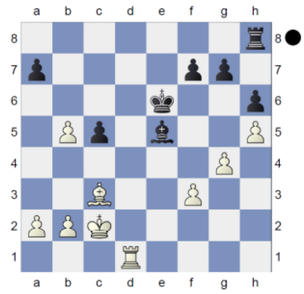
5



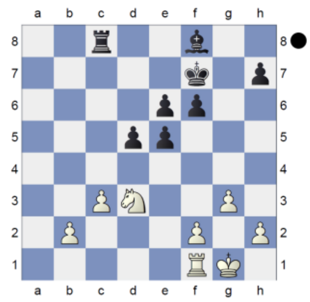
6



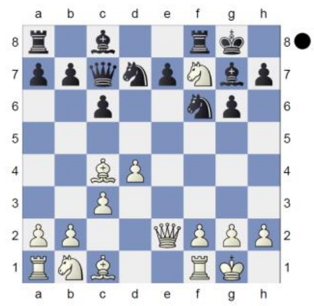
7



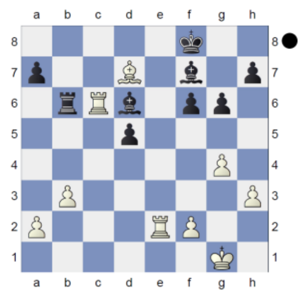
8



9



10



A1i

We will now provide you with some additional information about the ten positions.

We have asked **a player with a rating of 2335** to evaluate the ten games in the same conditions as you. You can find their prediction in the table below.

Looking back at your own evaluation in **Round 1, Part 1** on the Response Sheet, please complete **Round 1, Part 2**. You are free to change or keep your previous predictions based on the information on this sheet.

سنزودك الآن ببعض المعلومات الإضافية حول الاوضاع العشر.

لقد طلبنا من **لاعب(ة) تصنيفه 2335** أن يقيم الاوضاع العشر في نفس ظروفك. يمكنك الاطلاع على توقعاتهم في الجدول أدناه.

بالنظر إلى تقديرك في **الجولة 1، الجزء 1** في ورقة الإجابة، يرجى إكمال **الجولة 1، الجزء 2**. لك مطلق الحرية في تغيير توقعاتك السابقة أو الاحتفاظ بها بناءً على المعلومات الواردة في هذه الورقة.

رقم الوضع Position Number	التفوق Pawn advantage
1	-2.4
2	-0.7
3	-0.7
4	+2.4
5	+0.7
6	-0.7
7	+2.4
8	-0.7
9	-0.7
10	+2.4

You have a total of 4 minutes to complete this part.

لديك 4 دقائق لإكمال هذا الجزء.

A1ii

Now, we will repeat the previous exercise with a new set of ten positions.

Please complete **Round 2, Part 1** of the Response Sheet. This is the same procedure as for **Round 1.**

Once you have completed **Round 2, Part 1**, please wait for the experimenter to give you the next set of instructions.

الآن، سنكرر التمرين السابق بمجموعة جديدة من عشر أوضاع.

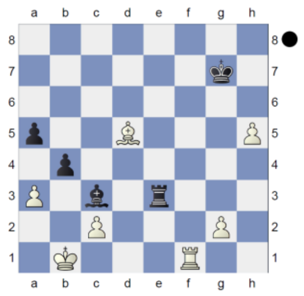
يرجى تعبئة **الجولة الثانية، الجزء 1** من ورقة الإجابة. هذا هو نفس الإجراء المتبع في الجولة الأولى.

بمجرد الانتهاء من الجولة 2، الجزء 1، من فضلك انتظر أن يعطيك المشرف المجموعة التالية من التعليمات.

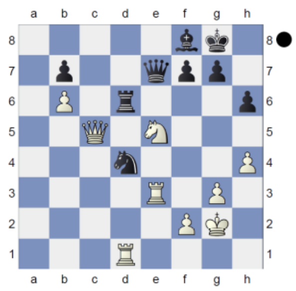
You have a total of <u>8 minutes</u> to complete this part.

لديك 8 دقائق لإكمال هذا الجزء.

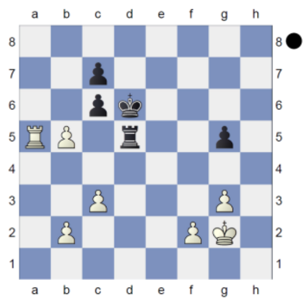
11



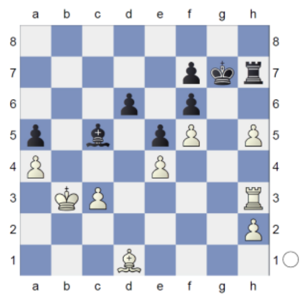
12



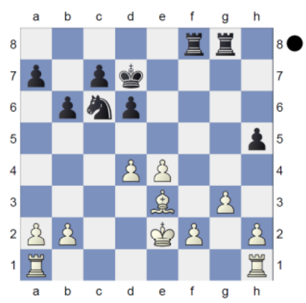
13



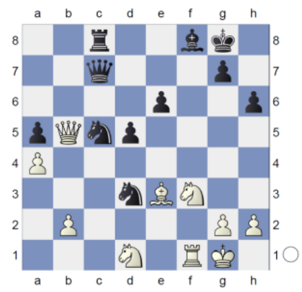
14



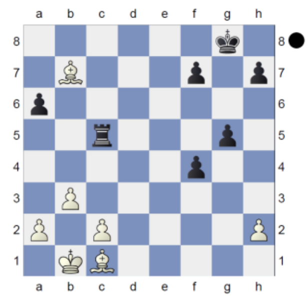
15



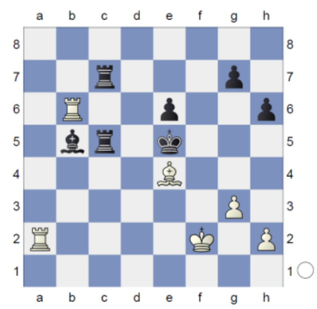
16



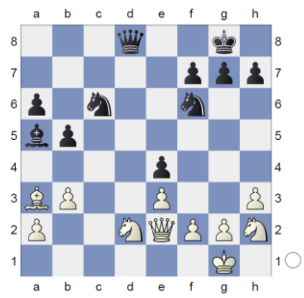
17



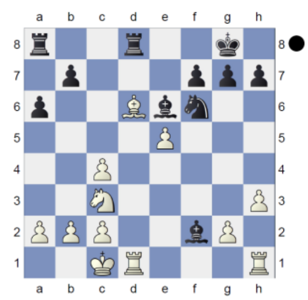
18



19



20



A1ii

We will now provide you with some additional information about the ten positions.

We have asked **an unrated player**, who plays regularly for fun, to evaluate the ten positions in the same conditions as you. You can find their predictions in the table below.

Looking back at your own evaluation in **Round 2, Part 1** on the Response Sheet, please complete **Round 2, Part 2**. You are free to change or keep your previous predictions based on the information on this sheet and to look at the prediction sheet.

سنزودك الآن ببعض المعلومات الإضافية حول الاوضاع العشر.

لقد طلبنا من **لاعب(ة)غير مصنف، يلعب بانتظام من أجل التسلية**، أن يقيم الاوضاع العشر في نفس ظروفك. يمكنك الاطلاع على توقعاتهم في الجدول أدناه.

بالنظر إلى تقديرك في **الجولة 2، الجزء 1** في ورقة الإجابة، يرجى إكمال **الجولة 2، الجزء 2**. لك مطلق الحرية في تغيير توقعاتك السابقة أو الاحتفاظ بها بناءً على المعلومات الواردة في هذه الورقة.

رقم الوضع Position Number	التفوق Pawn advantage
11	-2.4
12	+0.7
13	+2.4
14	-0.7
15	+0.7
16	-2.4
17	-0.7
18	-2.4
19	-0.7
20	-0.7

You have a total of 4 minutes to complete this part.

لديك 4 دقائق لإكمال هذا الجزء.

When this is over, please complete the personal information questions at the back of the response sheet.

عندما تنتهي من هذا الجزء، يرجى إكمال أسئلة المعلومات الشخصية في الجزء الخلفي من ورقة الإجابة.