Magic Mirror on the Wall, Who Is the Smartest One of All?

Online Appendix

Yoram Halevy Johannes C. Hoelzemann Terri Kneeland

January 18, 2025

A Experimental Results of All Participants

In this section, we replicate and report all results reported in the main text. Table A.1 presents the distribution of actions in the two diagnostic games.

Table A.1: Frequency of Action Choices in the Diagnostic Games

Action	IR	DS
a	298/470	36/470
b	63/470	82/470
c	59/470	352/470
d	50/470	—

All choices made irrespective of opponent type.

We begin by summarizing choice behavior and the preference relation over IR and DS irrespective of the opponent type. Table A.2 lists these results.

Table A.2: Preferences between IR and DS

	$IR \succ DS$	$IR \precsim DS$
IRM Prediction	*all*	*nil*
Ratio	212/470	258/470
Percentage	45.1%	54.9%

All choices made irrespective of opponent type.
$IRM \equiv$ Iterative 'top-down' model of reasoning.

As a next step, we control for participants whose behavior is inconsistent with best-responding across all games and either type. For example, we now remove participants who play a with a valuation $v > 12$, and further exclude those whose valuations exceed the maximum possible payoff given their action choice; e.g., playing b with a valuation $v > 13.25$ or c with a valuation $v > 12.25$ in either of the two control games, MS and NE. As a result, we are now focussing on 186 participants playing against an undergraduate student of any year or discipline and 180 participants playing against a Ph.D. students in Economics, respectively. Table A.3 lists these results of $n = 366$ choices irrespective of opponent type.

Table A.3: Controlling for Best-Response Consistency

	$IR \succ DS$	$IR \precsim DS$
IRM Prediction	*all*	*nil*
Ratio	166/366	200/366
Percentage	45.4%	54.6%

All choices made irrespective of opponent type excluding all choices that are inconsistent with best-responses in MS and NE. $IRM \equiv$ Iterative 'top-down' model of reasoning.

Next, we control for participants whose behavior is consistent with a preference for Nash equilibrium in pure strategies and either type. That is, we now remove participants who play c in both DS and NE as well as value this control game weakly above IR. This lets us focus on 173 participants playing against an undergraduate student of any year or discipline and 161 participants playing against a Ph.D. students in Economics, respectively. Table A.4 lists these results of $n = 318$ choices irrespective of opponent type.

Table A.4: Controlling for Nash Equilibrium Preference

	$IR \succ DS$	$IR \precsim DS$
IRM Prediction	*all*	*nil*
Ratio	163/334	171/334
Percentage	48.8%	51.2%

All choices made irrespective of opponent type excluding all choices that play c in DS and NE and value NE weakly above DS. $IRM \equiv$ Iterative 'top-down' model of reasoning.

Last, we leverage MS and NE and, in this step, exclude only those choices that value all small games equally; that is, $v_{DS} = v_{MS} = v_{NE}$. This results in concentrating on 137 participants playing against an undergraduate student and 126 participants playing against a Ph.D. students in Economics, respectively. Table A.5 lists these results.

Table A.5: Controlling for Equal Valuations of All Smaller Games

	$IR \succ DS$	$IR \precsim DS$
IRM Prediction	*all*	*nil*
Ratio	129/263	134/263
Percentage	49.0%	51.0%

All choices made irrespective of opponent type excluding all choices that value *DS*, *MS*, and *NE* equally.
IRM ≡ Iterative 'top-down' model of reasoning.

Overall, the inclusion of the controls does not alter the results. Similar to the results reported in the main text, while the ratio of those who weakly prefer *DS* over *IR* increases to some extent, using the entire sample also suggests that participants may value the predictability of their opponents' actions.

Turning to choices at the subject-level and a brief discussion of differences in behavior by opponent type. We have established that approximately half of the choices made by these participants are consistent with difficulty of predicting others' behavior. On the full sample, this turns out to be even stronger when we control for valuing all smaller games equally as highlighted above. Table A.6 shows the comparative statics of the ranking over the set of diagnostic games conditional on the opponent's identity (i.e., either an undergraduate student or a Ph.D. student in Economics).

Table A.6: Ranking of *IR* and *DS* by Opponent Type

			Undergraduate	
			$IR \succ DS$	$IR \precsim DS$
Ph.D.	$IR \succ DS$	*IRM* Prediction	*all*	*nil*
		Ratio	67/235	49/235
		Percentage	28.5%	20.9%
	$IR \precsim DS$	*IRM* Prediction	*nil*	*nil*
		Ratio	29/235	90/235
		Percentage	12.3%	38.3%

IRM ≡ Iterative 'top-down' model of reasoning.

Lastly, we ran ordinary least-square regressions with random effects controlling for order effects as well as the opponent order. In particular, we regressed the difference in valuations of *IR* and *DS* ($v_{IR} - v_{DS}$) on the opponent dummy *PhD*, which is 0 for facing an undergraduate student and 1 for playing against a Ph.D. student in Economics, and the valuations for both MS and NE. Further, we include the game order dummy *DS before IR*, which is 0 if IR is displayed before *DS* and 1 if *DS* is displayed before IR. In addition, we also include the opponent order dummy *PhD before UG*, which is 0 if participants played first against an undergraduate student and afterwards against a Ph.D. student in Economics in the first part of the experiment and 1 if the order is reversed.

Table A.7: OLS Estimations with Random Effects of Difference in Valuations of IR and *DS*

Ranking by *Opponent*	UG: $IR \succ DS$ PhD: $IR \succ DS$	UG: $IR \precsim DS$ PhD: $IR \succ DS$	UG: $IR \succ DS$ PhD: $IR \precsim DS$	UG: $IR \precsim DS$ PhD: $IR \precsim DS$	*All*
	$v_{IR} - v_{DS}$	$v_{IR} - v_{DS}$	$v_{IR} - v_{DS}$	$v_{IR} - v_{DS}$	$v_{IR} - v_{DS}$
Intercept	2.474^{***}	-1.075	2.498^{*}	-1.597^{**}	0.069
	(0.831)	(1.101)	(1.379)	(0.685)	(0.682)
PhD	-0.190	3.642^{***}	-3.418^{***}	0.206	0.360^{*}
	(0.186)	(0.290)	(0.350)	(0.148)	(0.170)
v_{MS}	-0.116	-0.043	0.007	0.037	-0.039
	(0.076)	(0.090)	(0.111)	(0.054)	(0.055)
v_{NE}	0.070	0.019	-0.007	0.030	0.067
	(0.078)	(0.094)	(0.115)	(0.057)	(0.058)
DS before IR					0.009
					(0.215)
PhD before UG					-0.225
					(0.219)
σ_{ϵ}	1.059	1.435	1.286	0.995	1.839
σ_u	0.897	0.750	0.812	0.961	1.002
N	134	98	58	180	470
(Between) R-squared	0.009	0.019	0.009	0.013	0.010

***Significant at the 1 percent level; **Significant at the 5 percent level; *Significant at the 10 percent level

We first split our sample by preference relation over the set of diagnostic games and opponent type (= 2 × 2) as in Table A.6 and then estimate the model using the full sample. Unlike in the main text, we do not exclude participants from our analysis whose valuations exceed the maximum possible payoff given their action and those who are inconsistent with best-responding in *DS*. Table A.7 lists the results from this analysis.

We find a strong effect of the observed characteristic of the opponent, *Ph.D.*, on the difference in valuations of IR and *DS* for all ranking as long as $DS \succsim IR$ against one opponent type only. This is also mildly true for the full sample, irrespective of the ranking over the set of diagnostic games. As expected, we do not find a strong effect when $DS \prec IR$. Here, we also do not observe a strong effect when $DS \succsim IR$. Overall, these estimation results for all $N = 235$ are in line with the difference in differences of valuations by opponent type and by ranking of *IR* and *DS* too. Using the full sample, we also do not find any indication of order effects, either due to presenting participants *IR* or *DS* before the other as well as playing each of the four games first against an undergraduate student or a Ph.D. student in Economics in the first part of the experiment.

B Further Analysis of Empirical Value Distributions

Moving beyond summary statistics, we now turn to the empirical distribution of valuations by the ranking of *IR* and *DS* induced by the valuations. We now enrich our discussion by leveraging the cardinal information obtained in the valuation task. Figure B.1 visualizes the empirical distributions of the valuations of the two diagnostic games, *IR* and *DS*, as well as the two control games, *MS* and *NE*. For this analysis we again focus on the 343 choices as summarized in Table 1 in the main text.

For the diagnostic games, the value distribution for *DS* (*IR*) is significantly higher (lower) in stochastic dominance when $DS \succsim IR$ than $DS \prec IR$: two-sample Kolmogorov-Smirnov test produces $p < 0.001$.[1] While differences between how the two "groups" value *IR* and *DS* are expected given how the groups are defined, the value distributions provide further support for the idea that the behavior of the $DS \succsim IR$ group refelcts reasoning that falls outside of the iterative 'top-down' model of reasoning. First, the large differences between the empirical value distributions in *IR* indicate that the $DS \succsim IR$ participants face difficulties in modeling and predicting the opponents' behavior in *IR* – a game where reasoning about rationality plays no predictive role. Second, participants' valuations in *DS* allows the analyst to infer their (confidence in their) beliefs about rationality: we can infer that participants with $12 \leq v \leq 12.25$ believe that their opponents are rational. Thus, the large difference between the empirical value distributions in *DS* indicates that the $DS \succsim IR$ group is more likely to believe in rationality relative to the

[1] In this discussion of empirical value distributions, all reported p-values are associated with two-sample Kolmogorov-Smirnov tests.

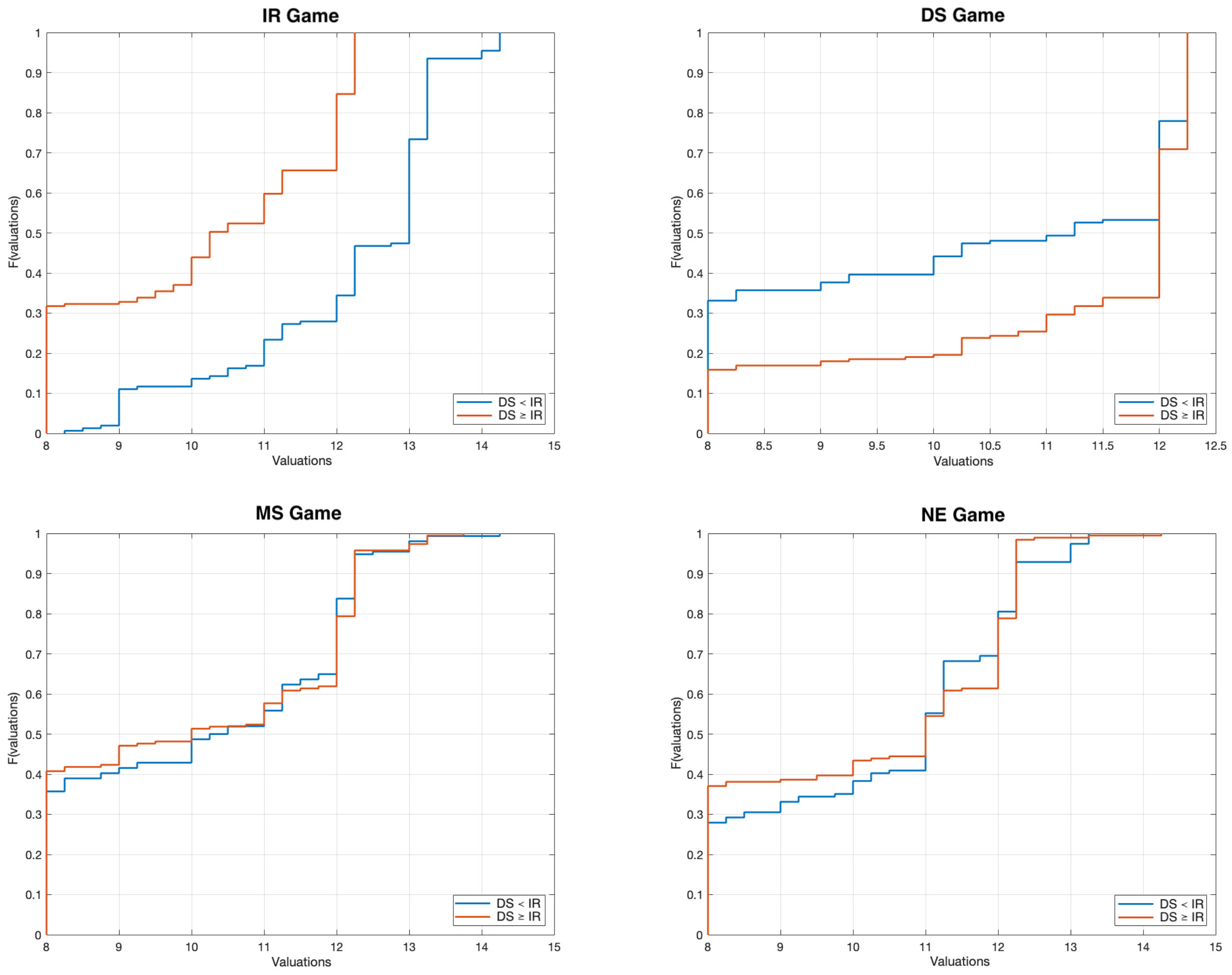
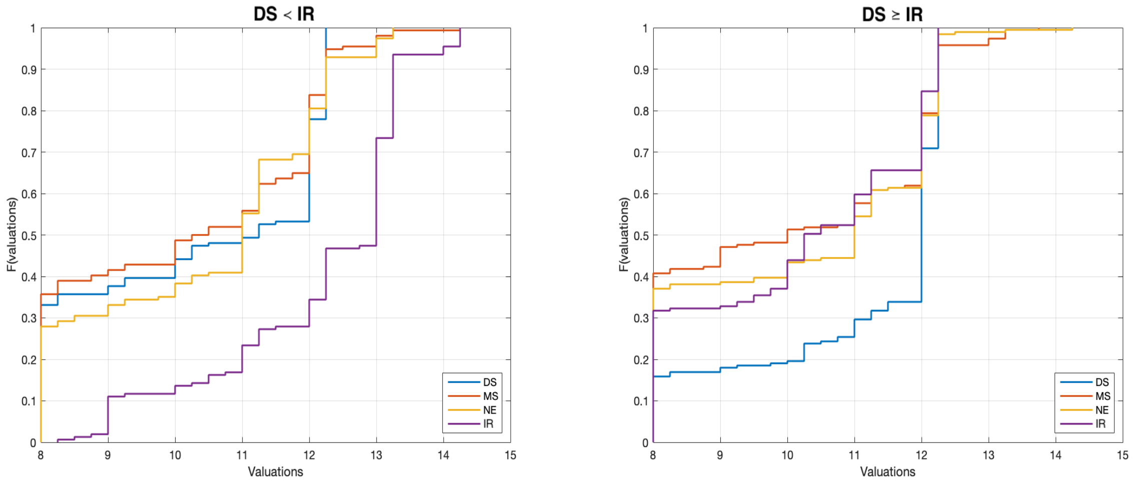


Figure B.1: Empirical Value Distributions of All Games by the Ranking of IR and DS for All $n = 343$ Choices. Top Row: The diagnostic games. Left: IR; Right: DS. Bottom Row: The control games. Left: MS; Right: NE.

$DS \prec IR$ group.

For the two control games, the empirical value distributions by ranking of IR and DS, the two groups of interest, overlap and cross each other several times as well. Thus, it is not surprising that no statistically significant differences can be detected ($p \geq 0.481$). This also supports the hypothesis that the relative preference for DS over IR between the two groups is not driven by a preference for small games or Nash equilibrium in pure strategies *per se* as these two groups value MS and NE similarly.

So far we only visualized the empirical value distributions separately for each game by the ranking of the set of diagnostic games. In Figure B.2, we show the empirical value distributions for all games by the ranking of IR and DS.



DS < IR

DS ≥ IR

Figure B.2: Empirical Value Distributions of IR, DS, MS, and NE by Ranking of IR and DS

For the $DS \succsim IR$ group, the valuation distribution for DS first-order stochastically dominates the valuation distributions of the two control games (both $p < 0.001$). Further, no statistical differences are observed when comparing the distributions of the two control games ($p = 0.429$). By contrast, when $DS \prec IR$, the valuation distributions of all small games overlap and are statistically indistinguishable from each other with the exception of DS and NE ($p = 0.035$).[2] We interpret these findings as further evidence that for approximately half of our participants, DS is indeed very attractive because it permits easier modeling and hence predicting the opponent's choices. The other half of participants, however, appear not to distinguish between the small games and, *inter alia*, have strictly higher valuations for IR than DS.

C Further Analysis of Opponent Type

By exploiting the cardinal information collected in the valuation task, we are able to detect not only ordinal differences in the ranking over the diagnostic games but also more nuanced differences: whether DS becomes *relatively* more or less attractive conditional

[2] Differences in valuation distributions are not significant: $p = 0.244$ from comparing games DS vs. MS and $p = 0.305$ for MS vs. NE, respectively.

on both the preference relation over *DS* and *IR* as well as the opponent's sophistication. The corresponding difference in differences of valuations $v_{IR} - v_{DS}$ by opponent type are depicted in Figure C.1.

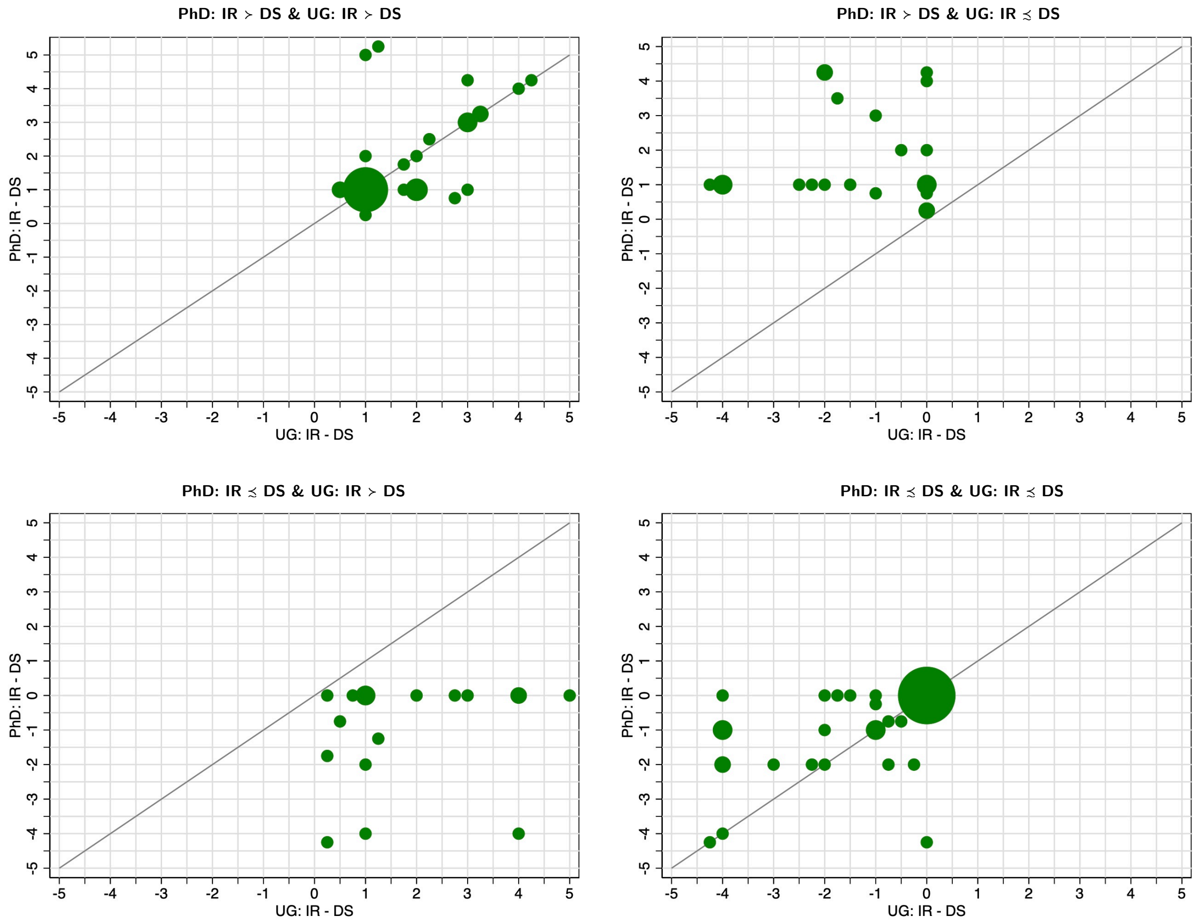


Figure C.1: Difference in Differences of Valuations of *IR* and *DS* by Ranking of *IR* and *DS and* by Opponent Type

As visualized in Figure C.1, depending on the preference relation over the games by opponent type, participants indeed value the games differently when facing either an undergraduate student or a Ph.D. student in Economics. On one hand, when $DS \succsim IR$ against both types, *DS* becomes relatively *less* valuable when playing against a Ph.D. student in Economics. This difference is statistically significant at the 5%-level using both t-test and Wilcoxon's signed-rank test ($p < 0.026$). On the other hand, when $DS \prec IR$ against both types of opponents, *DS* becomes relatively *more* valuable when facing a Ph.D. student in Economics. This difference, however, is not statistically significant ($p >$

0.257 for both tests). Naturally, whenever $DS \prec IR$ against one opponent type but not the other, the differences are statistically significant at the 1%-level (all $p < 0.001$). The direction of these asymmetries in the observed choices by opponent type surprised us. If anything, we conjectured *DS* becoming relatively *more* attractive when playing against a Ph.D. student in Economics conditional on ranking *DS* above *IR* (possibly because experiencing difficulties in predicting the opponent's choices).[3]

D Robustness Test

As a further robustness test and to complement the non-parametric analysis and key elements discussed in Section 4, we ran ordinary least-square regressions with random effects controlling for order effects as well as the opponent order. In particular, we regressed the difference in valuations of *IR* and *DS*, $v_{IR} - v_{DS}$, on the opponent dummy *PhD*, which is 0 when facing an undergraduate student and 1 when playing against a Ph.D. student in Economics, and the valuations for both *MS* and *NE*. Further, we include the game order dummy *DS before IR*, which is 0 if *IR* is displayed before *DS* and 1 if *DS* is shown before *IR*. In addition, we also include the opponent order dummy *PhD before UG*, which is 0 if participants played first against an undergraduate student and afterwards against a Ph.D. student in Economics in the first part of the experiment and 1 if the order is reversed.

To account for the fact that we observe each participant repeatedly and behavior across games for the same participant is not independent, we treat each participant as our units of statistically independent observations. We first split our sample by preference relation over the set of diagnostic games and opponent type (= 2 × 2) as in Table **??** and then estimate the model using the full sample. As above, we exclude participants from our analysis whose valuations exceed the maximum possible payoff given their action,

[3] The findings do not qualitatively change when we restrict attention to those participants who hold the belief that their opponent is rational. When *DS* is ranked above *IR* against both types, *DS* still becomes relatively *less* enticing when playing against a Ph.D. student in Economics. This difference is statistically significant at the 5%-level using both t-test and Wilcoxon's signed-rank test ($p < 0.034$). When *DS* is ranked below *IR*, *DS* still becomes relatively more alluring when facing a Ph.D. student. It is not statistically significant ($p > 0.160$ for both tests), as in the aggregate-choice analysis. As above, when *DS* is ranked above *IR* against one opponent type but not the other, the differences are also statistically significant at the 1%-level (all $p < 0.008$).

those who played any other action than c in DS, and those who are inconsistent with best-responding in MS and NE.[4] Table D.1 lists the results from this analysis.

Table D.1: OLS Estimations with Random Effects of Difference in Valuations of IR and DS

Ranking by	UG: $IR \succ DS$	UG: $IR \precsim DS$	UG: $IR \succ IR$	UG: $IR \precsim DS$	*All*
Opponent	PhD: $IR \succ DS$	PhD: $IR \succ DS$	PhD: $IR \precsim DS$	PhD: $IR \precsim DS$	
	$v_{IR} - v_{DS}$	$v_{IR} - v_{DS}$	$v_{IR} - v_{DS}$	$v_{IR} - v_{DS}$	$v_{IR} - v_{DS}$
Intercept	2.571^{***}	-0.743	2.772	-1.566^{*}	0.246
	(0.933)	(1.338)	(1.742)	(0.925)	(0.866)
PhD	-0.038	3.308^{***}	-2.620^{***}	0.357^{**}	0.291^{*}
	(0.135)	(0.378)	(0.502)	(0.179)	(0.173)
v_{MS}	-0.050	-0.119	-0.216	0.079	-0.071
	(0.091)	(0.111)	(0.174)	(0.065)	(0.067)
v_{NE}	-0.018	0.046	0.105	-0.025	0.073
	(0.088)	(0.119)	(0.160)	(0.076)	(0.073)
DS before IR					-0.030
					(0.277)
PhD before UG					-0.197
					(0.281)
σ_{ϵ}	0.619	1.276	1.141	0.884	1.375
σ_{u}	1.241	0.549	1.215	1.025	1.471
N	96	53	33	109	291
(Between) R-squared	0.030	0.514	0.426	0.031	0.012

***Significant at the 1 percent level; **Significant at the 5 percent level; *Significant at the 10 percent level

We find a strong effect of the observed characteristic of the opponent, *Ph.D.*, on the difference in valuations of IR and DS for all ranking as long as $DS \succsim IR$ against at least one opponent type. This is also mildly true for the full sample, irrespective of the ranking over the set of diagnostic games. As expected, we do not find a strong effect of type when $DS \prec IR$. These estimation results are in line with the difference in differences of valuations by opponent type and by ranking of IR and DS, as depicted in Figure C.1. We do not find any indication of order effects, either due to presenting participants IR or DS before the other as well as playing each of the four games first against an undergraduate student or a Ph.D. student in Economics in the first part of the experiment.

[4] We replicated the same analysis on the entire sample and report the results in the Online Appendix.

E Detailed Non-Choice Data Analysis

In this section, we report detailed results that were only concisely presented in the main text in Section 4.4. As text data required more data cleaning and preprocessing, we performed the following steps. For normalization, we converted the data to a consistent format, e.g., lowercasing. Next, in terms of tokenization we split text into words, phrases, symbols, or other meaningful elements. Further, we removed common words that may not add value to the analysis, i.e., stop word removal. In addition, we reduced words to their base or root form, that is, stemming or lemmatization. Lastly, in order to handle special characters and punctuation, we removed or replaced non-alphanumeric characters as necessary.

E.1 Exploratory Data Analysis

In order to identify the most common words or phrases, we begin with a simple and straightforward frequency analysis. The top ten most common words across the entire dataset, excluding common English stop words are "player" (200 occurrences), followed by "choose (198), "highest" (113), "12" (66), "option" (64), "best" (56), "action" (49), "earnings" (48), "pick" (48), and "row" (47), respectively.

Next we turn to length analysis, which involves examining the distribution of text lengths across our dataset to gain insights into the structure and the nature of the text by ranking over the two diagnostic games and for each game separately. Figure E.1 visualizes the implementation of the two diagnostic games. It appears that participants tend to write more detailed comments, measured by average word and sentence count, about their reasoning in games that they prefer. For example, participants who rank *IR* above *DS* write, on average, 35.03 (1.5) words (sentences) in *IR* but only 29.33 (1.14) words (sentences) in *DS*. By contrast, those who rank *DS* above *IR* write 31.53 (1.33) words (sentences) in *DS* and just 30.25 (0.97) words (sentences) in *IR*.

We move on to visualize key terms and their frequencies as word clouds in Figure E.2.

In the next step of our exploratory analysis, we focus on differences in participants'

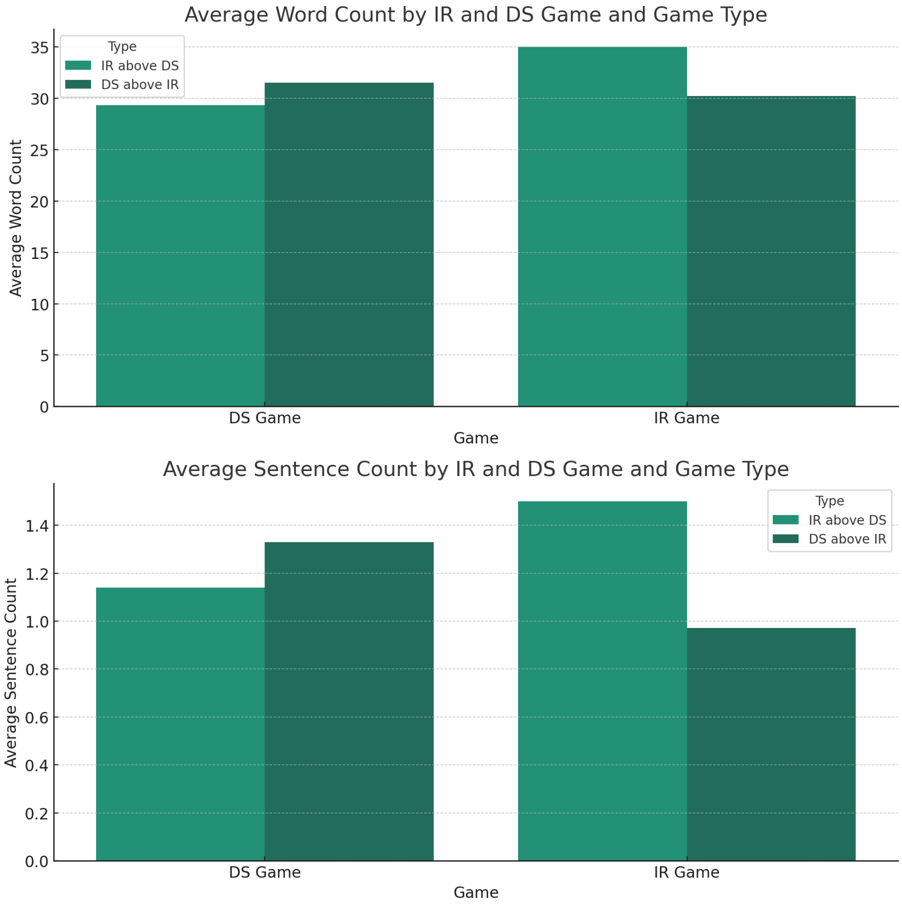


Figure E.1: Average Word Count (top) and Average Sentence Count (bottom)

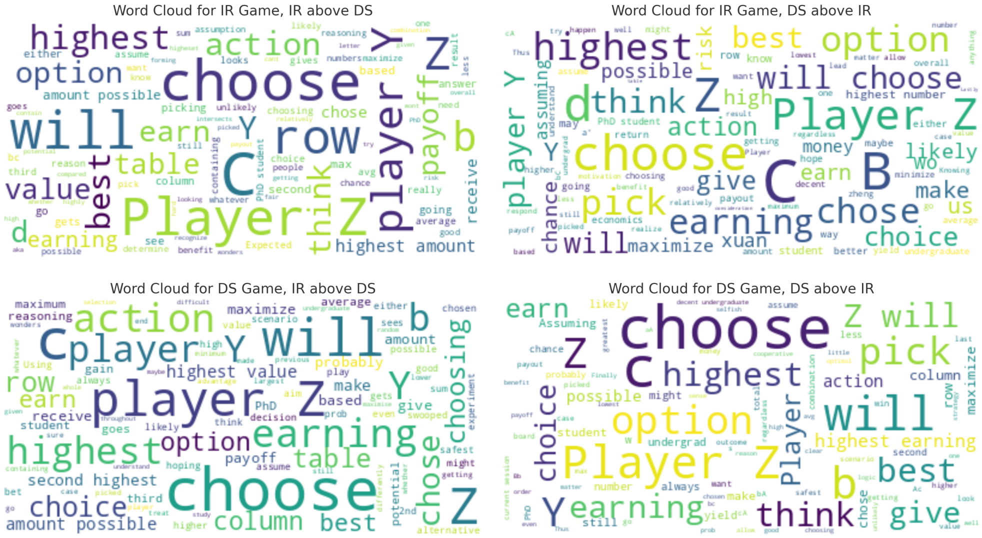


Figure E.2: Word Clouds by the Ranking of *IR* and *DS*. Top Row: *IR* Game; Bottom Row: *DS* Game. Left Column: $IR \succ DS$; Right Column: $IR \precsim DS$.

notes. In particular, we highlight the unique words most commonly used within each ranking over the games. Figure E.3 illustrates these unique keywords by ranking and for each game separately.

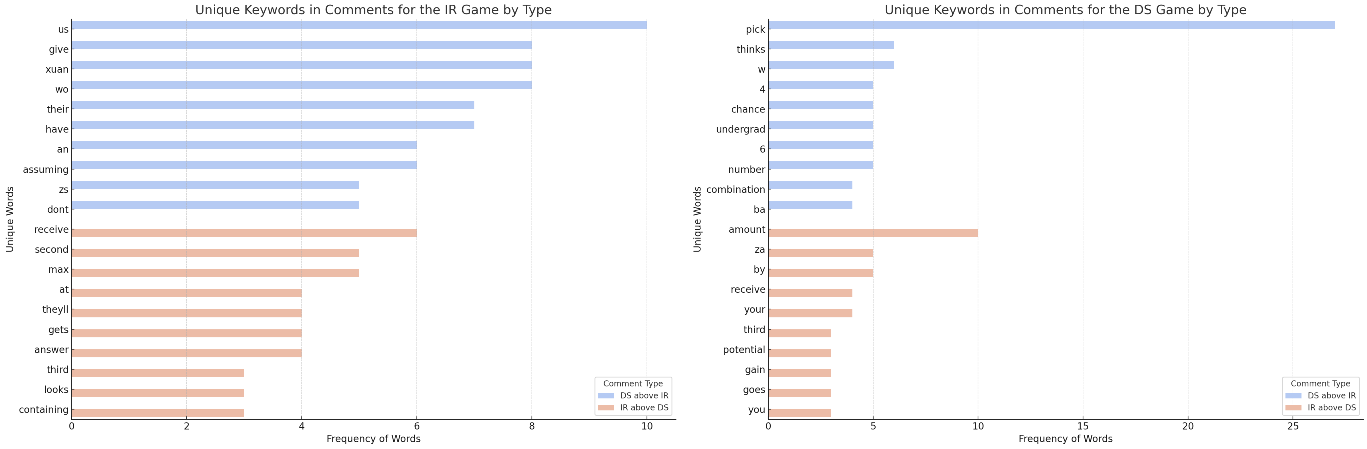


Figure E.3: Unique Keywords Used by the Ranking of *IR* and *DS*. Left Column: *IR* Game; Right Column: *DS* Game.

Before we conclude our exploratory analysis, we delve into complexity indicators. As we have seen in Figure E.1, participants' ranking over the two diagnostic games, as inferred by their choices, is associated with higher average word and sentence counts. The frequency of complexity-related keywords within notes written could serve as a proxy for participants' ability to express more complex reasoning processes in the diagnostic game that they rank above the other. Here, we focus on two specific measures that can serve as proxies for the complexity discussed: complexity keyword frequency and average comment length. First, the frequency of predefined complexity-related keywords within participants' notes can serve as a direct indicator of a strategic complexity discussion. Higher frequencies of these keywords may suggest more in-depth strategic considerations. The complexity keywords used in the analysis are terms that hint at strategic thinking, decision-making processes, and considerations of various options or outcomes. Examples of such keywords are “strateg,” “decid,” “choos,” “option,“ “think,” “consider,” “outcome,” “possibl,” or “predict.” Second, longer comments might indicate more elaborate discussions, potentially reflecting the ability to express higher strategic complexity. The average note length for each ranking over *DS* and *IR* can thus serve as a proxy for the level of detail and complexity in the discussions. Figure E.4 illustrates these two complexity measures.

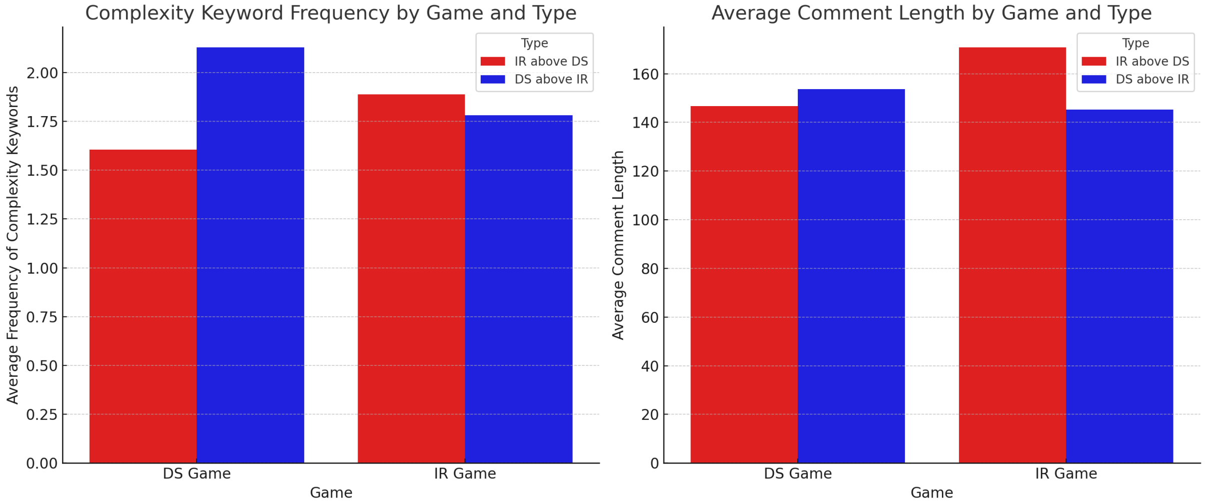


Figure E.4: Complexity Measures by Ranking of IR and *DS* and IR and *DS* Game. Left: Complexity Keyword Frequency; Right: Average Length of Notes Taken.

In *DS*, notes made by those who rank *DS* above IR tend to include more complexity-related keywords and are slightly longer on average compared to notes taken by participants who rank *DS* below IR. This is suggestive evidence that discussions involving those who prefer *DS* over IR might delve deeper into strategic deliberation when it comes to predicting behavior in *DS*. In IR, however, both ranking types show a higher frequency of complexity keywords compared to *DS*, with those who rank IR above *DS* notes being significantly longer on average. This is suggestive evidence that IR prompts more complex strategic deliberations, especially for $IR \succ DS$, where the discussions are not only more frequent in terms of complexity-related keywords but also more detailed, as indicated by the longer comment length. Overall, these findings suggest that the strategic complexity discussed in participants' notes varies by both diagnostic game and ranking over the games, with discussions in *DS* by those who rank *DS* above IR and discussions in IR by those who rank IR above *DS* exhibiting higher levels of complexity, as indicated by both the frequency of complexity-related keywords and the average comment length.

E.2 Feature Extraction

We now proceed with feature extraction such as Bag-of-Words (BoW) to represent the notes to "their future-self" as a matrix of token counts; Term-Frequency-Inverse-Document-Frequency (TF-IDF) to reflect the importance of a term to a comment relative to the overall corpus; as well as Word-Embeddings and thus use pre-trained vectors like Word2Vec and GloVe to capture semantic meanings of words. In Figure E.5, we highlight and visualize the word embeddings for words found in our dataset, projected into two dimensions using principal component analysis (PCA) for ease of visualization. Each point represents a word, and its position in the space is determined by the PCA transformation of the document-term matrix, simulating how words might be represented in a high-dimensional embedding space.

This serves as a visual approximation of word relationships based on their occurrence across notes written by participants. Words that are closer together in this two-dimensional space are more likely to have similar contexts within the dataset. By contrast, words that are further apart are less related.e22

E.3 Modeling and Analysis

Let us now turn to more elaborate modeling and techniques. We begin with topic analysis on participants' notes and use Latent-Dirichlet-Allocation (LDA), a popular method for topic modeling. This approach allows us to identify distinct topics present in the notes and to understand the distribution of these topics across the two games and rankings over the games.

In turn, we examine what topics are most relevant or correlate with participants' ranking over the two diagnostic games and the two games of interest, respectively. To do so, we study the distribution of topics within each note to participants' "future self" and then aggregate this information by ranking and game. We assign the most dominant topic to each note based on the LDA model output and compute the proportion of each topic within each type–game combination.[5] Figure E.6 visualizes the topic distribution

[5] In this section of the Appendix, we use the terms "type" and "ranking over the games" interchangeably.

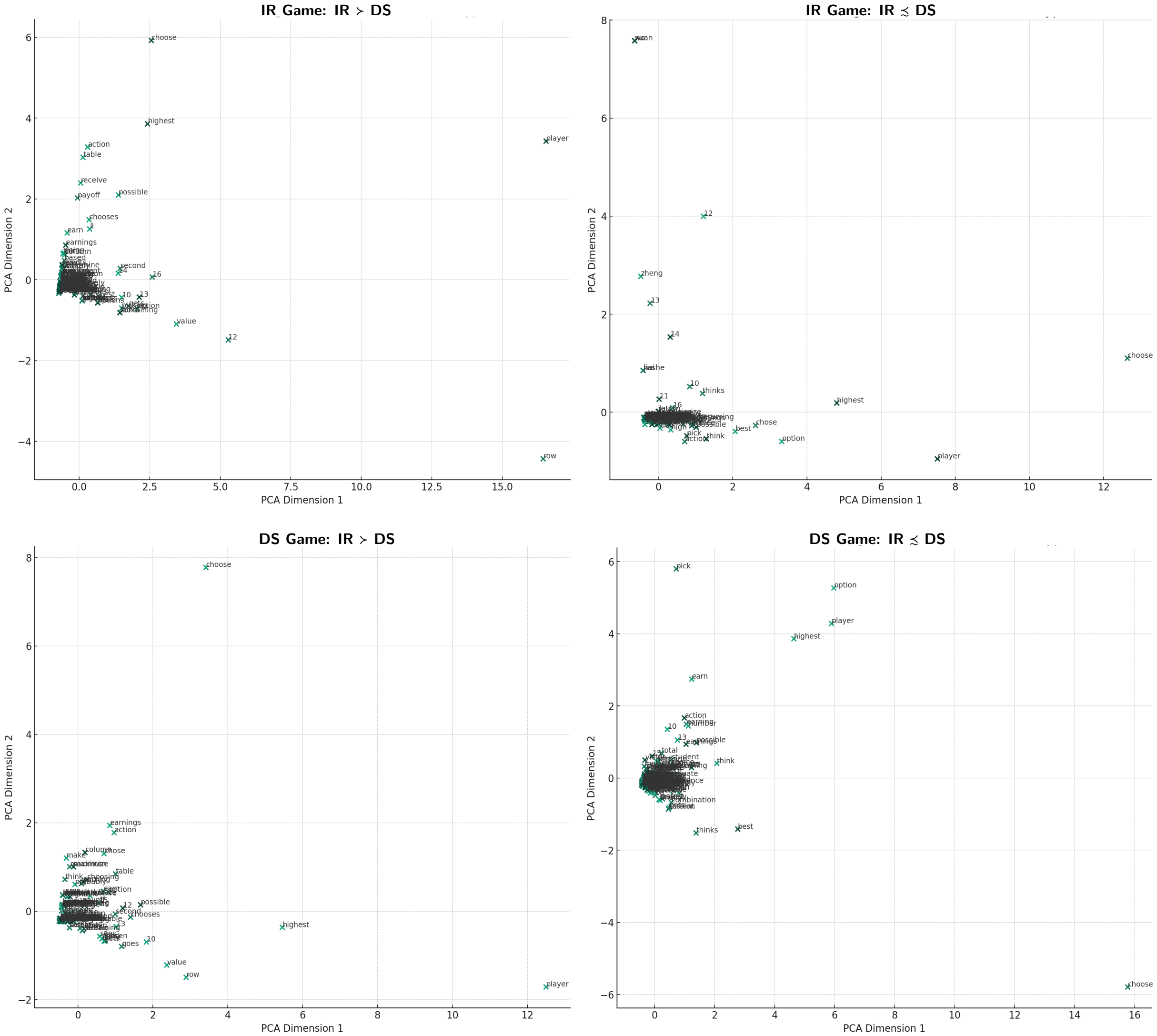


Figure E.5: Simulated Word Embeddings by the Ranking of *IR* and *DS*. Top Row: *IR* Game; Bottom Row: *DS* Game. Left Column: *IR* ≻ *DS*; Right Column: *IR* ≾ *DS*.

of the two diagnostic games by participants' ranking over these.

These proportions indicate qualitative evidence that a higher emphasis on Topic 3 (in both games) is associated with ranking *DS* above *IR*, while ranking *IR* above *DS* is associated with more emphasis on Topic 2 in *DS* and Topics 4 and 5 in *IR*.

In the next step, we focus on sentiment analysis to determine the sentiment expressed in the notes, in particular, whether participants are more positive, negative, or neutral in their expressions. Average sentiment polarities by ranking over the two diagnostic games differ significantly. For those who rank *DS* above *IR*, the average sentiment polarity is approximately 0.162 while those participants who prefer *IR* over *DS* display an average

Table E.1: Topic Analysis Using Latent Dirichlet Allocation

Topic	Keywords	Interpretation
1	choose, option, think, player, best, highest, ll, possible, thinks, going	Seems to be about making decisions or choices, considering the best or highest options available.
2	12, choice, 13, 10, 15, 11, action, earnings, choices, ca	Appears to focus on numerical aspects or quantitative choices, potentially related to specific actions or earnings.
3	player, highest, chose, option, choose, pick, best, earning, earnings, make	Similar to Topic 1, this topic also revolves around decision-making, focusing on choosing the best or highest earning options.
4	player, row, highest, choose, action, best, 12, possible, 14, second	Could be discussing strategies involving rows or positions, with a focus on choosing the best or highest-ranking actions.
5	earn, pick, earning, player, choose, highest, column, earnings, maximize, max	Centered around maximizing earnings or benefits, with emphasis on picking or choosing options that yield the highest earnings.

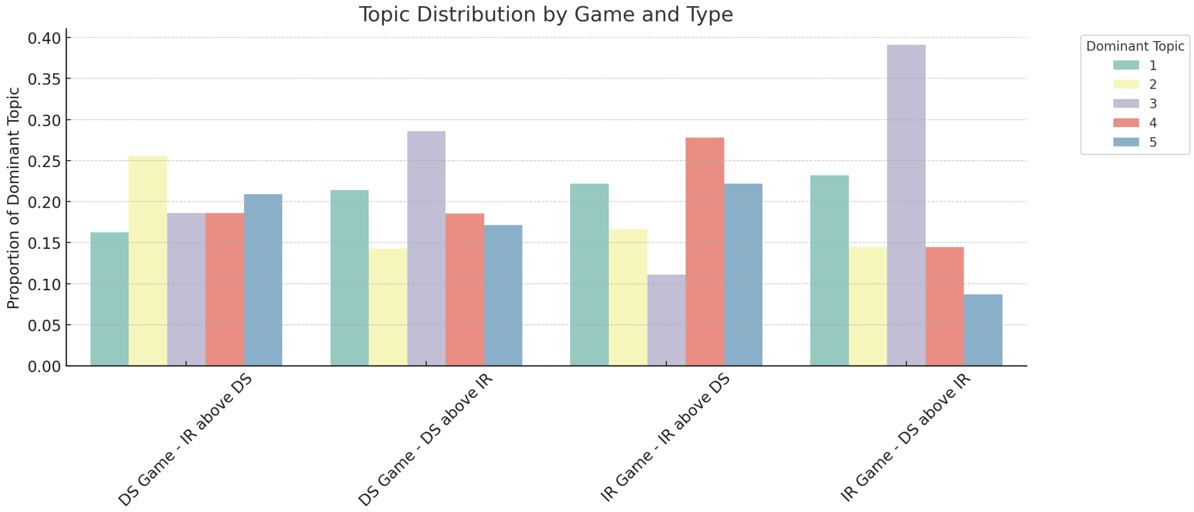


Figure E.6: Topic Distribution by Ranking of *IR* and *DS*. Games on the Left: *DS* Game; Games on the Right: *IR* Game.

sentiment polarity of roughly 0.128.

These results suggest that participants who rank *DS* above *IR*, on average, express

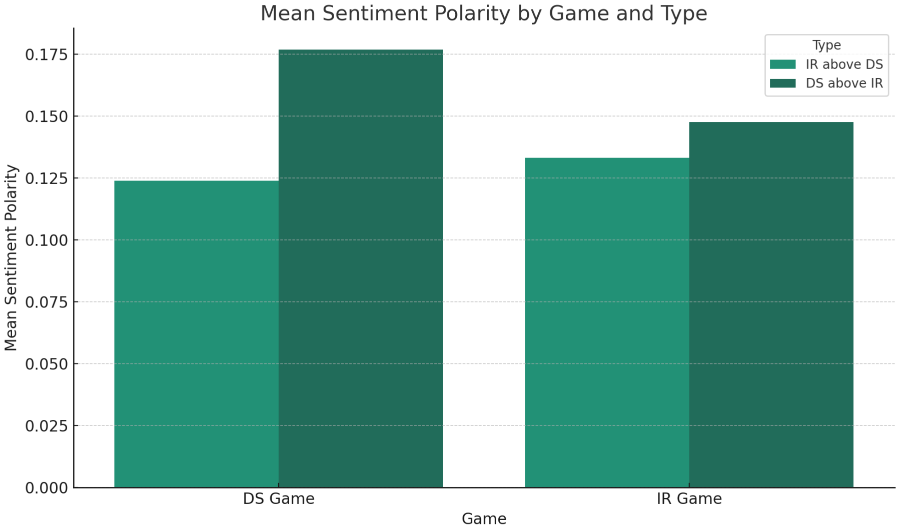


Figure E.7: Average Sentiment Polarity by Ranking of *IR* and *DS*. Left: *DS* Game; Right: *IR* Game.

comments with slightly more positive sentiment compared to those who prefer *IR* over *DS*. However, as Figure E.7 highlights and in line with participants ranking over the games, whenever *DS* is ranked above (below) *IR* the notes to their "future-self" indicate that they are also more positive (negative) in *DS* compared to *IR*.

We complement our sentiment analysis by analyzing the use of modal verbs that might indicate certainty or predictions in participants' notes to further explore confidence and prediction behavior. Figure E.8 illustrates the average certainty modal verbs count by ranking over the games and *DS* and *IR*, respectively.

The analysis of modal verbs that offers suggestive evidence of certainty or predictions shows that whenever a given participant ranks one diagnostic game over the other, then their choices are also associated with more certainty modal verbs per note written. For those who rank *DS* above *IR*, the average verbs count decreases from 0.914 to 0.478 when moving from *DS* to *IR*, suggesting a stronger confidence or a greater willingness to make firm predictions in *DS*. By contrast, participants who prefer *IR* over *DS* feature an increase in their average certainty modal verbs count from 0.698 in *DS* to 0.889 in *IR*, potentially indicating an increased confidence or predictive stance in *IR*.

Finally, we conclude our in-depth text analysis with a cluster analysis where we group texts based on similarity of content. We perform a cluster analysis on participants' notes,

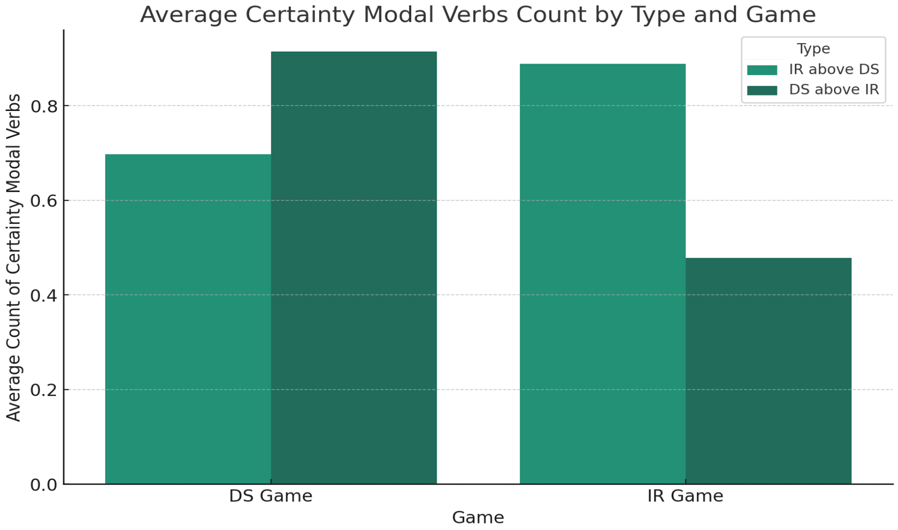


Figure E.8: Average Certainty Modal Verbs Count by Ranking of *IR* and *DS*. Left: *DS* Game; Right: *IR* Game.

use the document-term matrix (DTM), and apply a clustering algorithm to group participants' notes to their "future-self" based on their textual content. The common approach for clustering textual data that we follow here is the K-Means algorithm, which partitions the notes into clusters with similar word usage patterns. In a first step, we use both the elbow method and the silhouette score based on our dataset's characteristics to determine the appropriate number of clusters, eventually settling on five clusters.[6] Next, we apply the K-Means clustering algorithm to the DTM. To understand the content of each cluster identified, we offer here the most frequent and distinctive words in participants' notes belonging to each cluster. This involves analyzing the text data to identify keywords that are particularly representative of the comments within each cluster. These are summarized in Table E.2.

These keywords offer some qualitative insights into the thematic content of each cluster. While Clusters 1 and 4 seem to focus on numeric values and options, possibly related to strategic decisions or evaluations within games, other clusters like Cluster 2 emphasize decision-making with terms like "choose" and "chooses," alongside positional references like "highest" and "table." By contrast, Cluster 3 reflects contemplation and

[6] Details are available upon request.

Table E.2: Cluster Analysis Keywords

Cluster	Keywords
1	player, row, 12, value, gets, 10, option, 13, highest, 16
2	highest, player, choose, possible, option, table, chooses, chose, column, assuming
3	choose, player, think, best, will, earnings, thinks, option, highest, maximize
4	choose, 12, highest, pick, player, best, choice, 10, chose, earn
5	player, action, choose, earnings, highest, best, think, pick, chose, option

strategy with words like "think," "best," and "maximize," possibly indicating a focus on optimizing outcomes. Lastly, Cluster 5 mixes elements of decision-making like "choose" or "option" with an emphasis on outcomes as, e.g., "earnings" or "highest." Figure E.9 visualizes the discussions and considerations present within participants' notes, categorized by the clustering algorithm based on textual content similarities by ranking over the diagnostic games and for each of the games individually.

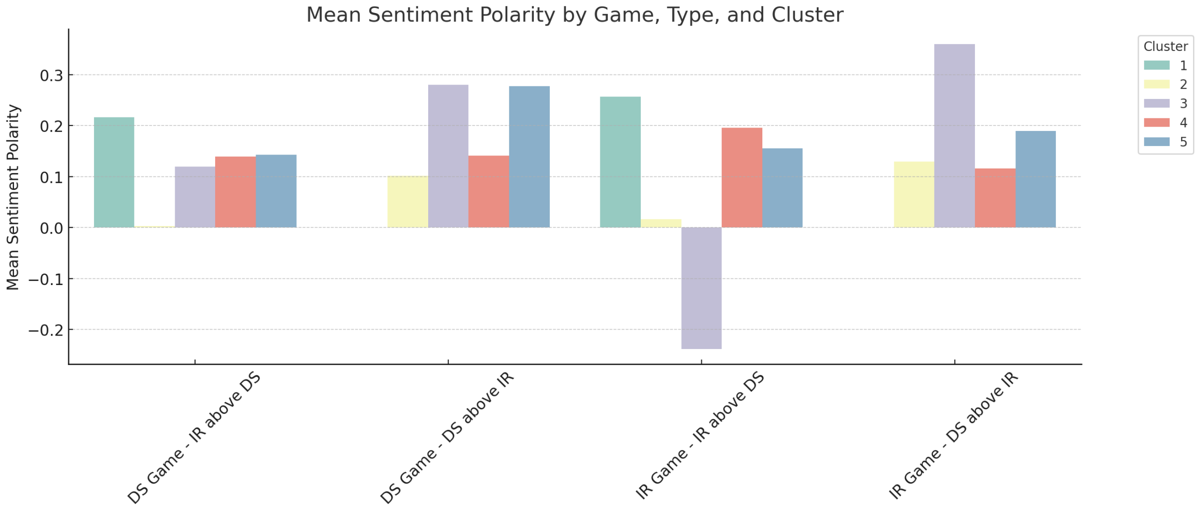


Figure E.9: Topic Distribution by Ranking of *IR* and *DS*. Games on the Left: *DS* Game; Games on the Right: *IR* Game.

As can be seen in Figure E.9, clusters are differently distributed across the two diagnostic games and across the ranking over the games. In particular, positive sentiment to Cluster 1 is associated with ranking *IR* above *DS*, while positive sentiment to Cluster 3 is associated with ranking *DS* above *IR*.