Case 2: An Analysis of the Chopin Competition

[redacted] Damary Gutierrez Hernandez, [redacted]

1 Introduction

1.1 Background Information

The Fryderyk Chopin International Piano Competition is the most important musical event in Poland and one of the most prestigious musical events in the world. Since its creation in 1927, the event has been organized every 5 years (with some delays or cancellations) and held at the Warsaw National Philharmonic. Competitors are tasked with playing a repertoire consisting of Fryderyk Chopin's works spanning his entire lifetime. The most recent competition–the 18th International Chopin Competition–was held in October 2021 and saw the participation of 151 pianists in the preliminary stage. In order to compete, participants must have been born between 1990 and 2004 and meet a series of requirements that include proof of superior piano-playing. Of the competitors who participated in the preliminary round, 78 made it to the main stage (Stage I), and an additional 9 pianists qualified directly by winning major piano competitions. Each of the subsequent stages consisted of fewer participants, with Stage II of the competition consisting of 45 participants, Stage III of 23, and the Final stage of 12. Each stage differed in programs, with programs consisting of a specific combination of études, nocturnes, ballades, scherzos, and mazurkas (among others) depending on the stage. Furthermore, the pieces could be performed in any order, with contestants prohibited from playing the same piece again in different stages of the competition (except for some pieces from the preliminary round). For the final program, contestants could select only one of two concertos ("XVIII Konkurs Chopinowski, Warszawa" n.d.).

Competitions of this nature are relatively popular; some other notable piano competitions include the Tchaikovsky International Music Competition, the Van Cliburn International Piano Competition, and the Queen Elisabeth International Music Competition ("Top Piano Competitions in the World" 2020). While these competitions go to great lengths to ensure that each participant is being judged fairly–the Chopin competition has two scoring systems, for example, not all competitions are fair in practice. An analysis by Flores and Ginsburgh of the ranking system of the Queen International Music Competition showed that final rank is not independent of performance order; those who appeared first had a statistically lower chance of being top-ranked, while those who performed on the last day had a higher chance of ranking high (Flôres and Ginsburgh 1996). Perhaps it is the case that repertoire type or order could have a similar impact on competition outcomes. Additionally, while teachers of contestants were not allowed to vote on their pupils, jurors do have access to those who vote and are in a position to feed their opinions onto other voters, and the competition jury contained several teachers of top contestants (Lebrecht 2021). Furthermore, variables such as piano choice, opus genre, and number of pieces played, for example, could also have a "repertoire effect," whether it was intended or unintentional from the judges perspective.

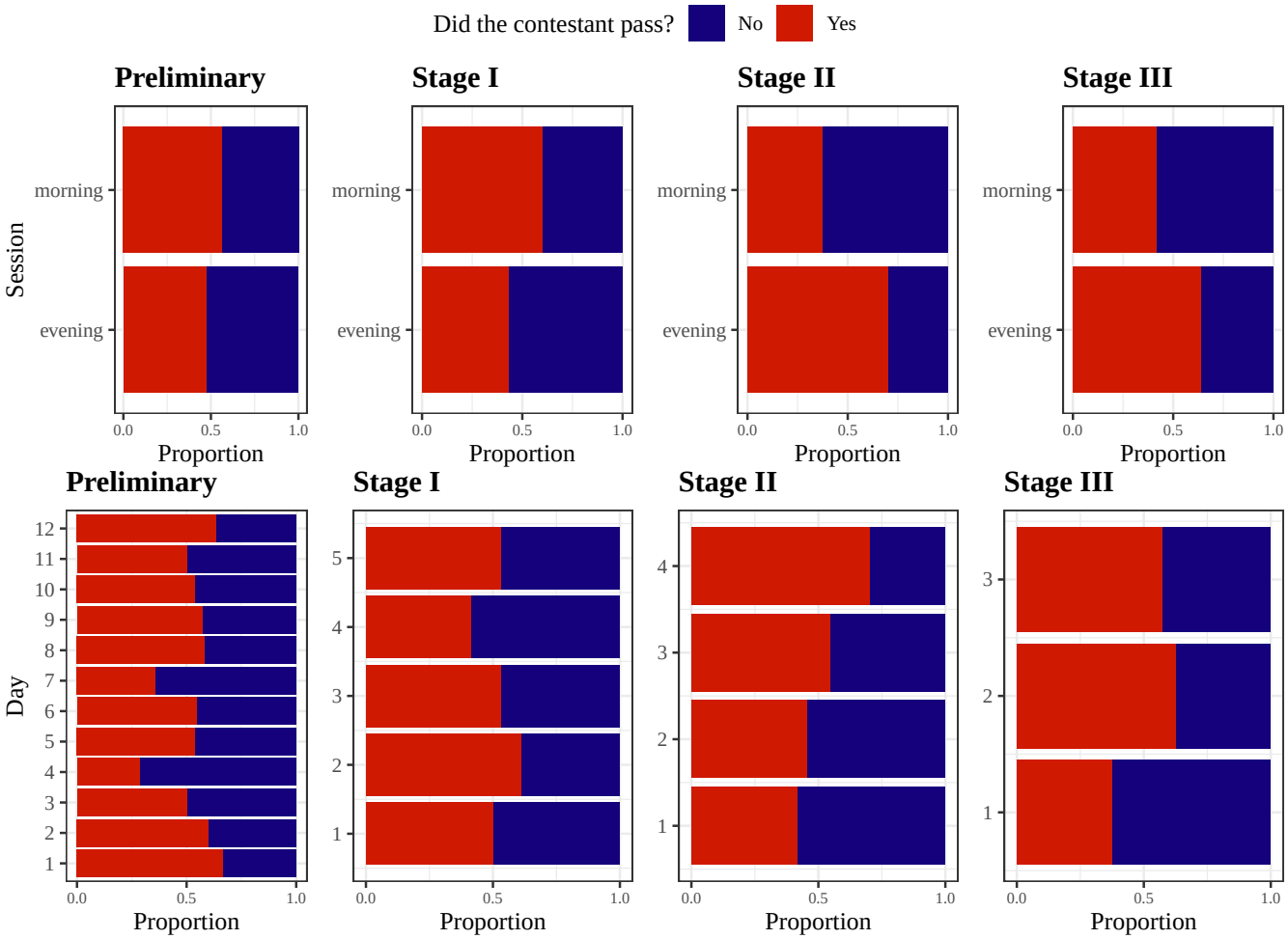
In this case study, our aims are twofold. We first aim to create a descriptive analysis of the entire competition by analyzing the trends found when focusing on "lucky" or "unlucky" contestants. "Unlucky" contestants are those who scored more points than the required threshold for Stages I-III and more than other contestants who did move on to the next stage but were themselves eliminated. In contrast, "lucky" contestants are those who also met the point threshold and moved on to the next stage, but scored below other contestants who did not move on. Next, we aim to evaluate whether there might be potential "repertoire effects" present in the competition. To do so, we will analyze whether certain repertoire choices are associated with better competition outcomes by creating a Bayesian generalized linear model for Stage I of the competition with the number of points gained as the response variable. The points variable is chosen as the dependent variable in order to account for lucky or unlucky contestants who performed well (or not so well) but didn't (or did) make it to the next stage of the competition due to the proportion of "Yes" verdicts.

1.2 Data Description

We are using a dataset containing the results and repertoire choices (in order) for each of the competitors in the 18th International Chopin Competition. Each participant's name, country, piano model used, and pieces played in each stage are included. Furthermore, variables describing whether a contestant passed each of the rounds are included, as are variables describing the proportion of "Yes" verdicts received from the jury and the number of points awarded for each stage. Lastly, the dataset also contains placements of those who made it to the Final. The dataset initially had 160 observations, each representing a competitor, and 38 variables (before the creation of new variables and data cleaning that our team conducted). Since piece selections were listed only as opus numbers in the original dataset, we matched each opus number with its genre (etude, nocturne, ballade, etc.) using information from the competition's Wikipedia page and included that as categorical variables ("XVIII International Chopin Piano

Competition" 2022). We also added the performance order of each of the contestants–as seen in the competition calendar–by creating variables that refer to what day the contestant performed and whether it was during the morning or evening session ("XVIII Konkurs Chopinowski, Warszawa" n.d.).

Figure 1. Initial analysis of day and time the contestant performed as a potential "repertoire effect"



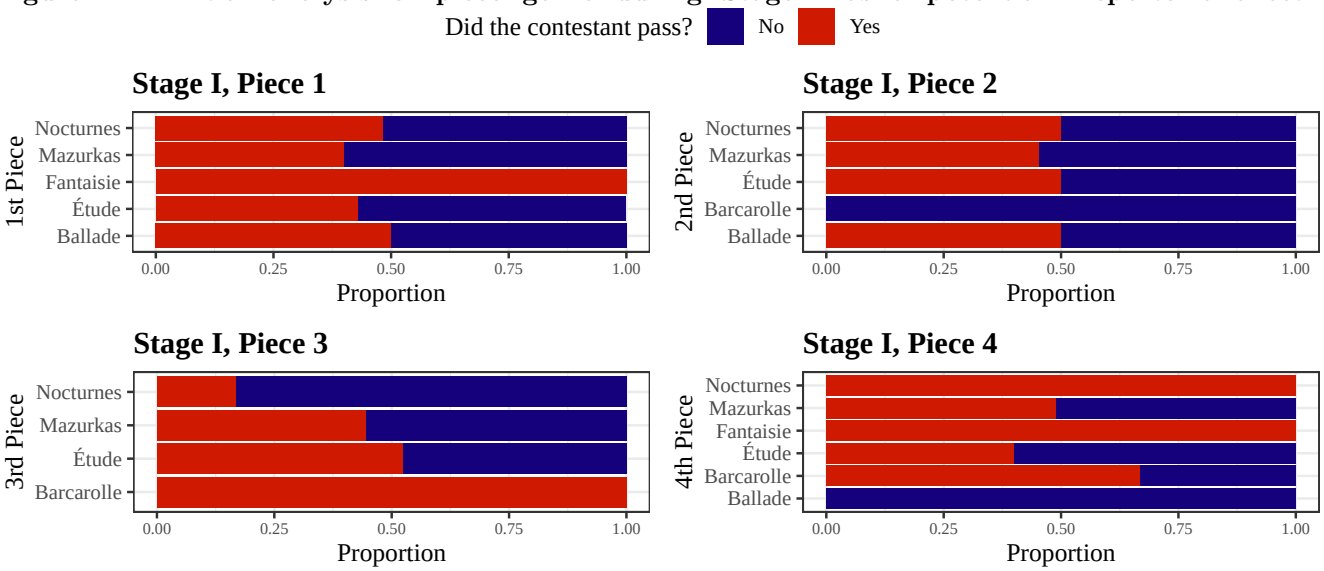
For an exploratory data analysis, we wanted to get a preliminary understanding of potential "repertoire effects" to develop a hypothesis and help drive our analysis. Each chart compares whether competitors passed each of the rounds with different repertoire choices to see if they could be associated with better competition outcomes.

When comparing the time at which the contestant performed (morning or evening session) to whether they passed to the next round, we see in Figure 1 that those who played in the morning session for the Preliminary and for Stage I were more likely to continue to the next round than those who played in the evening session. The opposite is true for Stages II and III with those who played in the evening session having a greater chance of making it to the next round when compared to those who played in the morning session.

While we see no clear trends by day in the Preliminary and Stage I, there does seem to be a trend during Stage II; the proportion of contestants who make it to Stage III increases the later the day. This is backed by Flores and Ginsburgh's study on music competitions, who predicted that the final rank is not independent of the day to which the candidate appears, as those who perform first have a lower chance of being ranked among the top (Flôres and Ginsburgh 1996). One possible explanation for this could include the "learning process" for judges in their evaluation, starting out with higher expectations and more strict rules and adapting their criteria eventually to the reality of the actual performances. As Flores mentioned in their study, it could be better to reassess earlier performances after the judges get a better idea of how all the competitors performed overall. "A similar phenomenon is at work when teachers grade written examinations: quite often, the first corrected papers are reassessed after a better idea of the performance of all students is gained." It is interesting to note as well that the winner of the competition, Bruce Liu, performed on the last day of the Final Stage, Stage III, Stage II, and Stage I. Futhermore,

while not as stark, we also see that the proportion of contestants who make it to the Final is greater for the last two days of Stage III performances than for the first day.

Figure 2. Initial analysis of piece genre during Stage I as a potential "repertoire effect"



As seen in Figure 2, when looking at piece genre we wanted to analyze whether the genre selected by each competitor could be associated with a better outcome in Stage I. As we hypothesized, there is a clear difference in the proportion of those who passed Stage I based on their piece genre, which may signify a "repertoire effect" in the competition.

1.3 Variables

For a full description of variables, see Appendix C. Of particular importance is the variable detailing the adjusted average score of each contestant, which will be used as a response variable in our analysis.

Along with the variables used for the response variables, for each competitor we have information on their nationality, piano model, and pieces played for each round. We created several new variables to generate what we felt were the most ideal predictors for our model. First, we created new variables for each piece that the competitor played to signify its genre. This variable returned an "NA" value if the competitor either did not advance to that round or performed less pieces in a particular round than fellow competitors. We also created variables for both the session time (either morning or evening) and day number that the contestant performed for each round. This was motivated by a study which stated that those who appeared first had a statistically lower chance of being top-ranked (Flôres and Ginsburgh 1996).

1.4 Hypothesis

We hypothesize that there will be "repertoire effects" that will be significantly associated with better competition outcomes in the 18th International Chopin Competition. We assert that performance order and piece genre (as well as other factors) will be correlated with better or worse outcomes for competitors.

2 Descriptive Analysis

In the 18th International Chopin Competition, in order to progress from one Stage onto the next, participants had to receive a majority of "Yes" votes from all members of the jury. In addition to these "Yes" or "No" votes, judges also had to assign participants a score for the round, and in order to assign a "Yes," this score had to be over a certain threshold which increased for each subsequent stage (from 17 points in the preliminary round to 20 points in Stage III; see Appendix C for more information). Both "lucky" and "unlucky" competitors were above this threshold for the respective stage. However, lucky competitors on average scored lower than other competitors who did not receive a majority of "Yes" votes whereas unlucky competitors had an average score for the stage which was above that of other competitors who did receive a majority of "Yes" votes, but they themselves did not garner a majority.

During the competition's three intermediate stages for which we have data on average points awarded to each competitor, there were 15 lucky and 14 unlucky competitors. The distribution across stages can be found in the table below. Overall, no competitor was ever lucky or unlucky more than once throughout the competition. Additionally, no competitor was ever lucky at one stage and unlucky at another stage during the competition. China had the

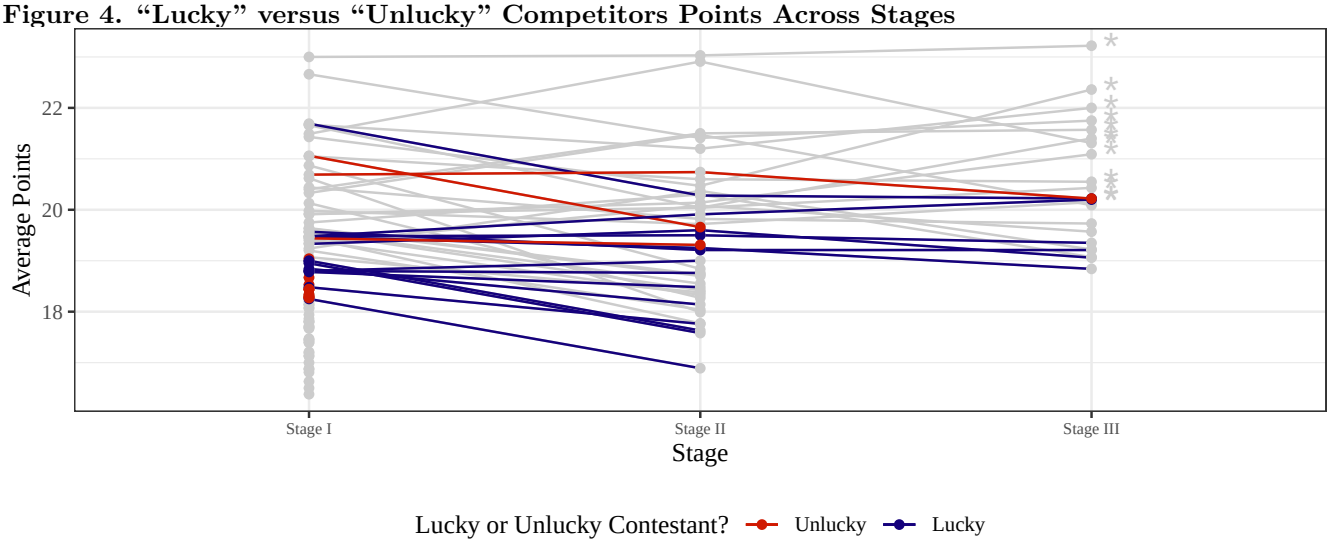
most unluckys with 8 competitors and also had no lucky competitors, while Poland had the most lucky competitors with 3 and 2 unlucky competitors.

Table 1: Lucky-Unlucky Distribution

Stage	Lucky	Unlucky
I	9	11
II	4	2
III	2	1

In the figure below, we can see that no competitor who got lucky in Stage I progressed past Stage II and no competitor who got lucky at Stage II progressed past Stage III (progression is marked by the presence of asterisks next to Stage III points). Additionally, the two competitors who got lucky in Stage III both were not awarded a standing after the final round. Finally, all three of the unlucky competitors in Stages II and III scored lower in the stage in which they were the unlucky than in the previous round. Note that in this figure, a single competitor was omitted due to their score of less than 15 points in Stage I.

Figure 4. "Lucky" versus "Unlucky" Competitors Points Across Stages



This lucky-unlucky phenomenon occurs in the competition because moving on to the next stage is not determined solely by points, but rather by obtaining a majority of judges who score you above the aforementioned minimum point threshold and who decide to vote "Yes" to move you on. Therefore, if a competitor has a small number of judges who score them much higher or much lower than the other judges, they may become an unlucky or lucky competitor. This effect of the average points reported is somewhat attenuated by the way in which high and low scores are added into the average for the competition. If a score is more than 3 points in Stage I or 2 points in Stages II-III above or below the true average, then the score will be modified to be exactly 3 points above or below this value and the average is recalculated. However, as we can see from the present analysis, the lucky-unlucky phenomenon still occurs even with this correction and serves to highlight the independence and equality each judge has in whether or not a participant will move on to the next stage.

3 Methodology

The goal of our analysis is to quantify and evaluate potential "repertoire effects" in the Chopin competition. Since the choice of repertoire is restricted to each stage and its respective program, we believe that it is more appropriate to focus on and model a particular stage of the competition. Furthermore, the number of contestants is filtered down at each successive stage of the competition so modeling later stages are constrained to small sample sizes. For example, 151 of the 160 total contestants participated in the preliminary round while only 12 contestants made it to the Final. In addition, the available outcome data differed between rounds. The preliminary round only had a binary outcome variable indicating whether a contestant passed or not. Meanwhile, Stages I to III had the same binary outcome variable along with the proportion of "Yes" responses from the jury and the average number of points awarded by the jury. In Stage I, a juror's "Yes" response is dependent on them awarding the contestant a score of 18 or higher. Therefore, the average number of points captures the overall evaluation of the jury and mostly captures whether a given contestant passed or not. Since Stage I has a larger sample size of 87 when compared to

later rounds while still offering the adjusted average points outcome, we believe that it is most appropriate to model the Stage I points outcomes to quantify and evaluate potential "repertoire effects" in the Chopin competition.

Using MCMC sampling, we fit a Bayesian generalized linear model on the adjusted average number of points scored by contestants in Stage I of the Chopin competition. As such, our coefficients are estimated β's which refer to posterior means. Also, our expected outcome is conditional. By sampling before fitting the model, we can work with a robust sample of representative data points that better approximates the true probability distribution of our data rather than our limited and observed distribution of Stage I points. All of the predictor variables are categorical variables indicating a contestant's choice of repertoire and the relative day and time they performed in the first round. By doing so, we believe our model captures potential "repertoire effects" on Stage I points while controlling for likely confounding factors on Stage I performances.

The MCMC sampling runs for 5 chains of 2000 draws each. Sampling ran on 1 core so it was done sequentially rather than in parallel. The first 1000 draws of each chain are for the warmup phase and the last 1000 draws of each chain are the posterior predictive samples for a total posterior predictive sample size of 5000.

Several model diagnostics suggest that our MCMC sampler has converged well. The plots and descriptions can be found in Appendix A. The gray posterior distributions in the p-p plots align fairly closely with the true distribution in blue which suggests that our sampler is generalizing the true data well. Our distribution of posterior predictions also aligns closely with the true distribution of Stage I adjusted average points which suggests that our sampler has converged and our model predicts well. The trace plot of each variable displays random scatter around the mean which suggests that the sampling chains have mixed well and converged. The $\hat{R}$ value of each variable is close to 1 which also suggests that our model has converged. While MCMC samples are dependent by nature, the high effective sample sizes suggest that our MCMC samples are only weakly correlated.

Below is our resulting Stage I points model in equation form:

$$\begin{aligned}\widehat{\text{Adjusted Average Points in Stage I}}_i =& 18.291 + 0.667 \times \text{I}(\text{First Choice}_i = \text{Etude}) \\ &+ 0.571 \times \text{I}(\text{Second Choice}_i = \text{Barcarolle}) \\ &- 0.104 \times \text{I}(\text{Second Choice}_i = \text{Fantaisie}) \\ &+ 0.717 \times \text{I}(\text{Second Choice}_i = \text{Scherzo}) \\ &+ 0.401 \times \text{I}(\text{Day}_i = 2) - 0.037 \times \text{I}(\text{Day}_i = 3) \\ &- 0.015 \times \text{I}(\text{Day}_i = 4) + 0.590 \times \text{I}(\text{Day}_i = 5) \\ &+ 0.159 \times \text{I}(\text{Time}_i = \text{morning})\end{aligned}$$

The first two predictor variables are categorical variables of the choice of repertoire in Stage I. Stage I required contestants to play four pieces with two of them coming from a list of etudes. However, the third piece required contestants to choose between 3 etudes and 6 nocturnes. The fourth piece required contestants to choose between 4 ballades, 1 barcarolle, 1 fantaisie, and 4 scherzos. Since the first two pieces draw upon the same set of Etudes in Op. 10 and Op. 25, we believe that the choice between different Etudes within the same piece is less significant in the outcome of Stage I points than the choice between different styles of pieces for the third and fourth piece. Therefore, our first predictor variable indicates whether a contestant chose an etude or a nocturne for their third piece and our second predictor variable indicates whether a contestant chose a ballade, barcarolle, fantaisie, or scherzo for their fourth piece. A contestant's four pieces can be played in any order so we will refer to a contestant's third piece choice as "First Choice" and their fourth piece choice as "Second Choice." We set nocturnes as the baseline category for "First Choice" and ballades as the baseline category for "Second Choice." These two predictor variables will help us evaluate and quantify potential "repertoire effects" in Stage I points in the Chopin competition.

The last two predictor variables are categorical variables for the day and time a contestant performed in Stage I. There are five day categories for the five days of performances during Stage I. We set day 1 as the baseline category. There are also two time categories for whether a contestant performed in the morning or evening session. We set the evening time as the baseline category. While not related to repertoire choices in Stage I, we believe that the day and time of a contestant's performance are the most likely confounding factors on the evaluation of Stage I performances. These two predictor variables will help us control for potential confounding factors on Stage I points in the Chopin competition.

Each juror can award a contestant a score between 1 to 25 inclusive. As mentioned above, a juror's "Yes" response in Stage I is dependent on whether they awarded a contestant a score of 18 or higher. This makes a score of 18 the cutoff point between passing and failing Stage I. The nature of the points scoring system motivated us to center our baseline around 18 to quantify how the choice of repertoire, day, and time nudged a given contestant towards

passing or failing. In order to achieve this, we set our prior intercept as a normal distribution with a mean of 18 and a standard deviation of 3 because the intercept of our model represents the baseline conditions of a contestant who played a nocturne as their "First Choice," played a ballade as their "Second Choice," and performed during the evening of day 1. By doing so, we are able to quantify how other choices of repertoire, day, and time are expected to raise or lower a contestant's score relative to the baseline. Since most contestants appeared to score at least in the range of 14 to 24, we believe that a standard deviation of 3 effectively focused the prior intercept distribution around these plausible values. However, we utilized the autoscaling feature of the `rstanarm` package so our prior intercept distribution's standard deviation was slightly increased to 4.571. Since we had little prior information on how the choice of repertoire, day, and time correlated with high or low performing Stage I outcomes, we believe that using the default priors for the coefficients–normal distributions with a mean of 0 and a standard deviation of 2.5–were appropriate as weakly informative priors. The default priors in `rstanarm` also utilized autoscaling so the adjusted standard deviations ranged between 7.580 and 13.925.

4 Results

Using the resulting model output, our model is as follows:

Table 2: Stage I Points Model

Variable	Mean	Mean SE	SD
(Intercept)	18.291	0.007	0.442
Etude 1st Choice	0.667	0.009	0.627
Barcarolle 2nd Choice	0.571	0.008	0.579
Fantaisie 2nd Choice	-0.104	0.007	0.490
Scherzo 2nd Choice	0.717	0.006	0.408
Day 2	0.401	0.008	0.529
Day 3	-0.037	0.009	0.549
Day 4	-0.015	0.009	0.541
Day 5	0.590	0.009	0.559
Morning Session	0.159	0.005	0.337
Sigma	1.545	0.002	0.125

Table 3: 95-Percentile Credible Intervals

Variable	Lower Bound	Upper Bound
(Intercept)	17.424	19.146
Etude 1st Choice	-0.575	1.888
Barcarolle 2nd Choice	-0.555	1.725
Fantaisie 2nd Choice	-1.063	0.886
Scherzo 2nd Choice	-0.068	1.546
Day 2	-0.616	1.414
Day 3	-1.105	1.041
Day 4	-1.019	1.078
Day 5	-0.508	1.685
Morning Session	-0.507	0.823
Sigma	1.318	1.815

In order to see if repertoire effects might be present, we can interpret the coefficients in our model pertaining to different repertoire choices. For example, a contestant who chose an etude between the two options in column three is expected to score 0.667 adjusted average points higher in Stage I than a contestant who chose a nocturne, holding all else constant, and has a 95% chance of scoring between 0.575 points lower to 1.888 points higher. Similarly, a contestant who chose a barcarolle between the four options in column four is expected to score 0.571 adjusted average points higher in Stage I when compared to a contestant who chose a ballade, with a 95% chance of scroing between 0.555 points lower to 1.725 points higher. In contrast, a contestant who chose a fantaisie for column four in Stage I is expected to score -0.104 adjusted average points higher than a contestant who chose a ballade, and has a 95% of scoring between 1.063 points lower and 0.886 points higher, holding all else constant. We also expect a 0.717 adjusted average point increase for a participant in Stage I who chose to perform a scherzo when compared to a participant who chose to perform a ballade, with a 95% chance they score between 0.068 points lower and 1.546 points higher.

While the day and time variables were added as controls, it is interesting to note that the coefficient for day two is 0.401, meaning that we expect a 0.401 adjusted point increase for a contestant who performs during day two of Stage I when compared to a contestant who plays on day 1, holding all else constant. We also expect to see a 0.590 adjusted point increase for participants who played on day 5 than on day 1. It could be the case that the jury is particularly hasher on day 1 of Stage I and that the performances of contestants who go last are the freshest in the jury's mind, leading to higher scores for those participants.

As evidenced by all of our 95% credible intervals crossing 0, none of our predictor variables were found to be significant at a 5% significance level. However, the intervals for etude as a first choice, barcarolle as a second choice, scherzo as a second choice, day 2, and day 5 suggest a higher concentration of plausible positive coefficient values than plausible negative coefficient values. In contrast, the intervals for fantaisie as a second choice, day 3, day 4, and morning session have similar concentrations of plausible positive and negative coefficient values. While none of our predictors may outright be deemed as significant, the higher concentration of plausible positive coefficient values in the 95% credible intervals for the previously mentioned variables provides evidence that these categories of the predictor variables may have some positive correlation with the adjusted average number of points for contestants in Stage I.

We must keep in mind the distribution of choice 1 and choice 2 piece types in our original data. For choice 1, 80 contestants chose to play a Nocturne while only 7 contestants chose to play an Etude. For choice 2, 42 contestants chose to play a ballade, 9 contestants chose to play the barcarolle, 14 contestants chose to play the fantaisie, and 22 contestants chose to play a scherzo. These uneven distributions may be confounding factors themselves. For example, the fact that the vast majority of contestants chose to play a nocturne instead of an etude suggests that the contestants themselves believed playing a nocturne is advantageous to playing a third etude. So the positive coefficient for etude as the second choice may be a result of the few contestants who chose to play a third etude performing at an extremely high level and impressing the jury with something unique while the larger number of contestants playing a nocturne averaged more mediocre performances and were more easily forgotten by the jury. Perhaps only 9 contestants attempting the barcarolle suggests that the barcarolle is a very difficult piece that only the top pianists at the competition are able to perform and excel at. In this case, the positive coefficient for barcarolle as the second choice may be a result of the few virtuoso pianists choosing to play the barcarolle who would have likely scored well regardless of their choice 2 piece. Overall, the uneven distributions of the two choice pieces provide evidence for both the contestants' self selection and the jury's biases, so we should be careful to keep this in mind in the discussion of our model coefficients. In our original sample size of 87, these uneven distributions would have been more problematic. By fitting our model on 5000 MCMC posterior samples, this alleviates some of our concerns with the limited sample size and uneven distributions.

5 Discussion

5.1 Conclusions

Our initial hypotheses going into this modeling exercise were that there would be significant associations between competitor order and piece genre with competition outcomes. We also hypothesized that these "repertoire effects" determined by our model would allow us to describe selections, that if made by competitors, would be associated with higher scores in the Chopin Competition, specifically in the first stage upon which our model was built. These predicted effects on score would be informative on a participant's competition outcome during Stage I since the score is the primary factor considered by judges when determining whether or not to vote to advance a competitor to the next stage of the competition.

Using Bayesian modeling techniques, we were unable to determine conclusively strong evidence for piece genre nor day of performance or time of day effects on a competitor's score in Stage I. Additionally, other variables relating to the competitors which we were initially interested in investigating likewise proved to be lacking evidence for significance and were omitted from our final model due to their inclusion resulting in poor model performance when it came to MCMC mixing. By not including these variables on piano type played and country of origin we potentially lose information on judge scoring bias which was introduced by these factors; however, by the reasoning above, we were confident that their exclusion was justified given the data we had available.

The most compelling repertoire effect occurs in the category in which participants were able to select between a ballade, barcarolle, fantaisie, or scherzo in Stage I. For this variable, while the 95% credible intervals do contain both negative and positive values, a choice of a Scherzo has a moderate chance of being positively correlated with points scored in Stage I.

This lack of strong evidence cannot in and of itself be used to suggest a definitive lack of correlation, and additional limitations which may lead to suboptimal model fit are discussed below. We can conclude that given the present data and our assumptions upon it which are incorporated in our model schema, we were unable to find strong evidence that our factors of interest, including the "repertoire effect" of piece selection, were correlated with score differentials in Stage I, with the exception of some evidence for scherzo selection with a positive correlation.

5.2 Limitations and Future Directions

Regarding modeling, although we categorized each piece by genre to analyze specific genres' association with competition outcomes, we did not look at which specific pieces within those genres are associated with better outcomes. More specifically, we did not look at specific etude pieces (i.e., Op. 10, No. 1 or Op. 25, No. 6), we only looked at the broader categories of the piece. Likewise, the main keys of different pieces and the tempo of the piece may also be used as a control for "repertoire effect" associated with better outcomes. Finally, the sample size of our model is relatively small and, while we use MCMC sampling to address this issue, it is still a limitation in our study. An additional limitation in our model is using points received as the response variable due to the analysis of "lucky" and "unlucky" competitors mentioned in our descriptive analysis. While we justified the use of points rather than a logistic model (with the response as whether the competitor passed to the next stage) as we wanted

to differentiate between those who barely got passed into the next round versus competitors who received very high or very low scores, the limitation in this analysis was that there were competitors who did not get passed that had higher average points in a stage than those who did make it through the round (and vice versa).

Currently, our analysis is done at the level of a single year of the competition and does not account for some other variables that could affect outcomes. With data from previous years of the competition, our model would be up to par with current literature analyzing potential "repertoire effects" of piano competitions instead of being isolated to a single competition in one year. Additionally, our current literature highlights that several contestants had teachers who were judges in the competition. While teachers were not allowed to vote on their pupils, they do have access to other jurors and therefore could have an effect on the outcome of the competitor (Lebrecht 2021). To account for this in the future, we can add a variable dedicated to contestants whose teachers are a part of the jury to assess if that is another potential "repertoire effect" that we did not account for. Another confounding factor in the competition is that contestants who have performed well in previous years or in other piano competitions could be more likely to perform better in the 2021 Chopin Competition, leading to an important bias potentially affecting the results. A variable tied to results from previous competitions could help control for the biases that go along with previous year performances. This could also be true for pianists who have participated in the competition previously, as a contestant who is seasoned and well-versed in the competition could be more likely to perform better than a first-year "amateur."

Finally, in the exploratory data analysis (Appendix B) we examined whether "home-field advantage" could be correlated with better results in the competition. The audience could potentially have an effect on the judges' decisions and they would often be more likely to cheer for contestants from Poland than pianists from outside the country.

Given all these additional perspectives, exploring these avenues of analysis would bolster our study as it would increase its scope, highlight more confounding factors, and provide a better sense of the overall "repertoire effects" that could be associated with better outcomes.

5.3 Summary

Overall, this study was unable to provide strong statistical evidence of "repertoire effects" that are associated with score differential in Stage I of the competition. While this lack of strong evidence could be deemed a disappointing outcome and a result of model limitations, it led to several meaningful conclusions, shedding light on variables that could be positively correlated with better competition outcomes and the potential lack of overall "repertoire effects" present in the competition. The model suggests that specific piece genres, for example, could potentially be positively correlated with a better chance of moving into Stage II, despite the credible intervals containing both negative and positive values. On the other hand, none of the variables that we hypothesized would be associated with better competition outcomes, including piece genre and competition order, provided strong statistical evidence of a being a "repertoire effect." There is little doubt that competition outcome may have a significant impact on the future course and career of the pianists competing, therefore it is essential that judging biases are minimized and "repertoire effects" relating to competition performance are limited.

6 References

Flôres, Renato G., and Victor A. Ginsburgh. 1996. "The Queen Elisabeth Musical Competition: How Fair Is the Final Ranking?" *Journal of the Royal Statistical Society. Series D (The Statistician)* 45 (1): 97–104. https://doi.org/10.2307/2348415.

Lebrecht, Norman. 2021. "Lessons to Be Learned from the Chopin Competition." Slipped Disc. October 21, 2021. https://slippedisc.com/2021/10/lessons-to-be-learned-from-the-chopin-competition/.

"Top Piano Competitions in the World." 2020. KeytarHQ: Music Gear Reviews. May 8, 2020. https://www.keytarhq.com/top-piano-competitions-in-the-world.html.

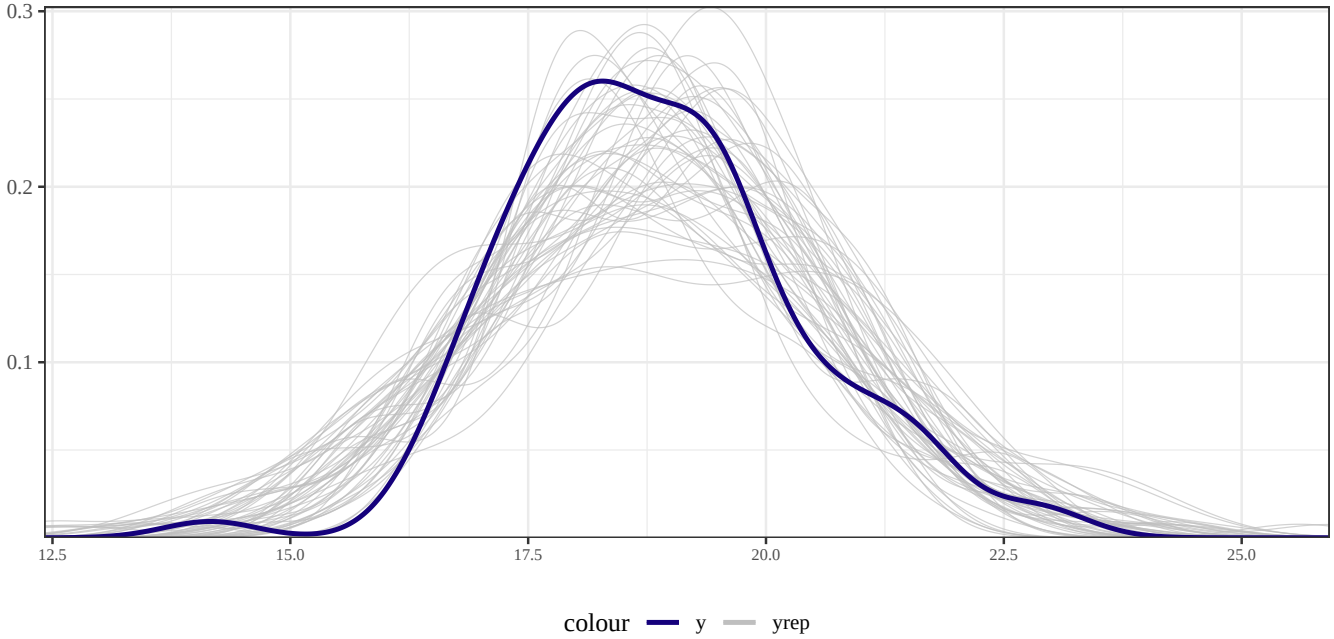
"XVIII International Chopin Piano Competition." 2022. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=XVIII_International_Chopin_Piano_Competition&oldid=1069841500.

"XVIII Konkurs Chopinowski, Warszawa." n.d. Accessed February 19, 2022. https://chopin2020.pl/en/.

7 Appendix

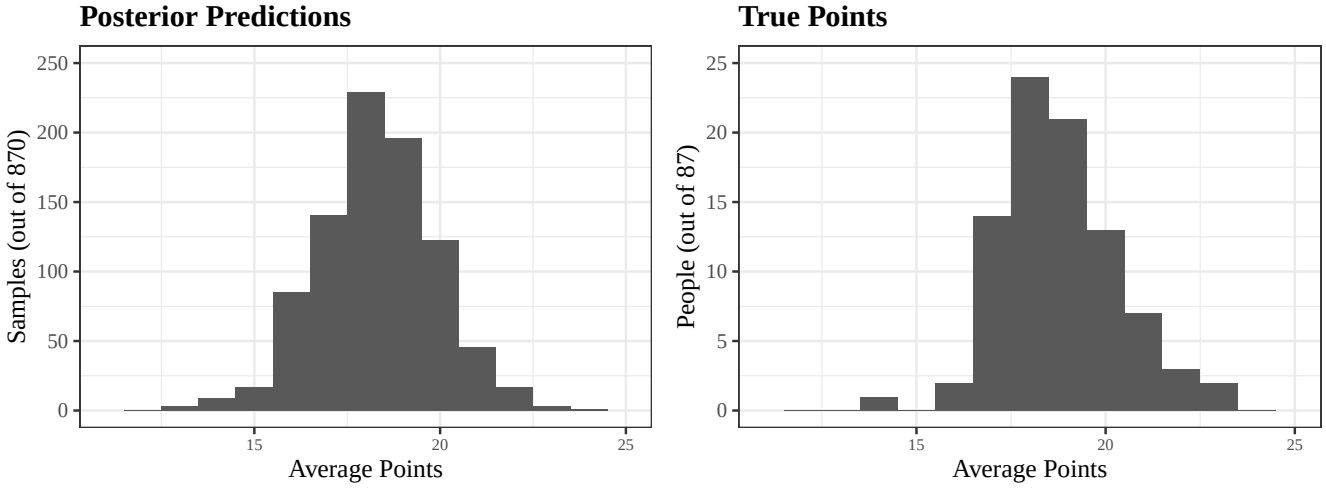
Appendix A: Model Diagnostics

A.1: P-P Plot



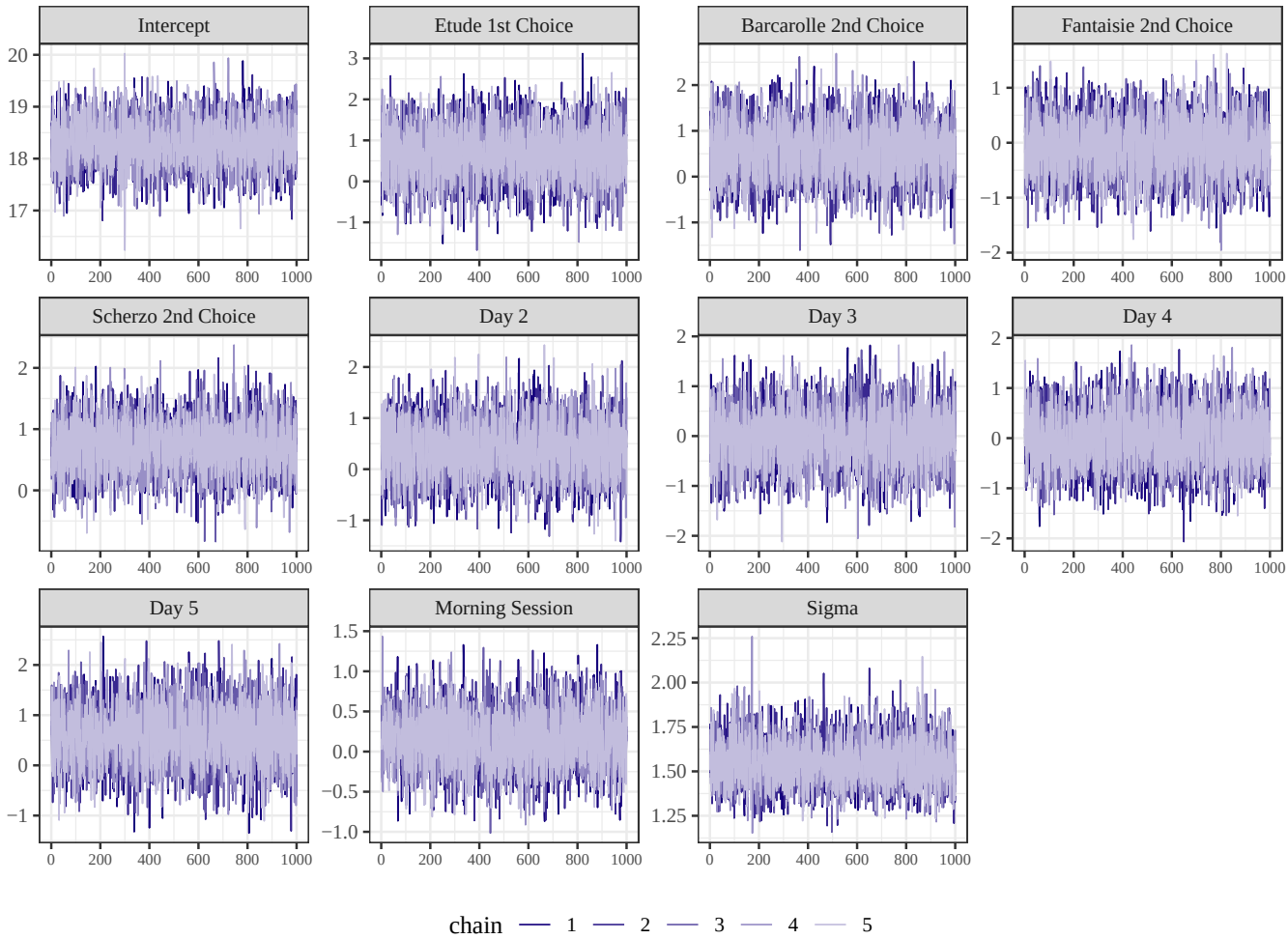
Above is the P-P Plot displaying our data's true distribution of Stage I adjusted average points in dark blue and our MCMC posterior distributions in gray. Apart from some outliers, the posterior distributions align fairly closely with the true distribution, suggesting that our model is able to generalize and evaluate potential "repertoire effects" as well as confounding factors on the Stage I adjusted average points.

A.2: Posterior Predictions



On the left is the distribution of 870 of our model's posterior predictions of Stage I adjusted average points. On the right is the distribution of the actual 87 contestants and their true Stage I adjusted average points. As evidenced by our posterior predictions distribution aligning closely with our true points distribution, our model appears to fairly accurately predict the expected number of adjusted average points in Stage I when given a contestant's first choice piece, second choice piece, day of their performance, and time of their performance.

A.3: Trace Plot Convergence



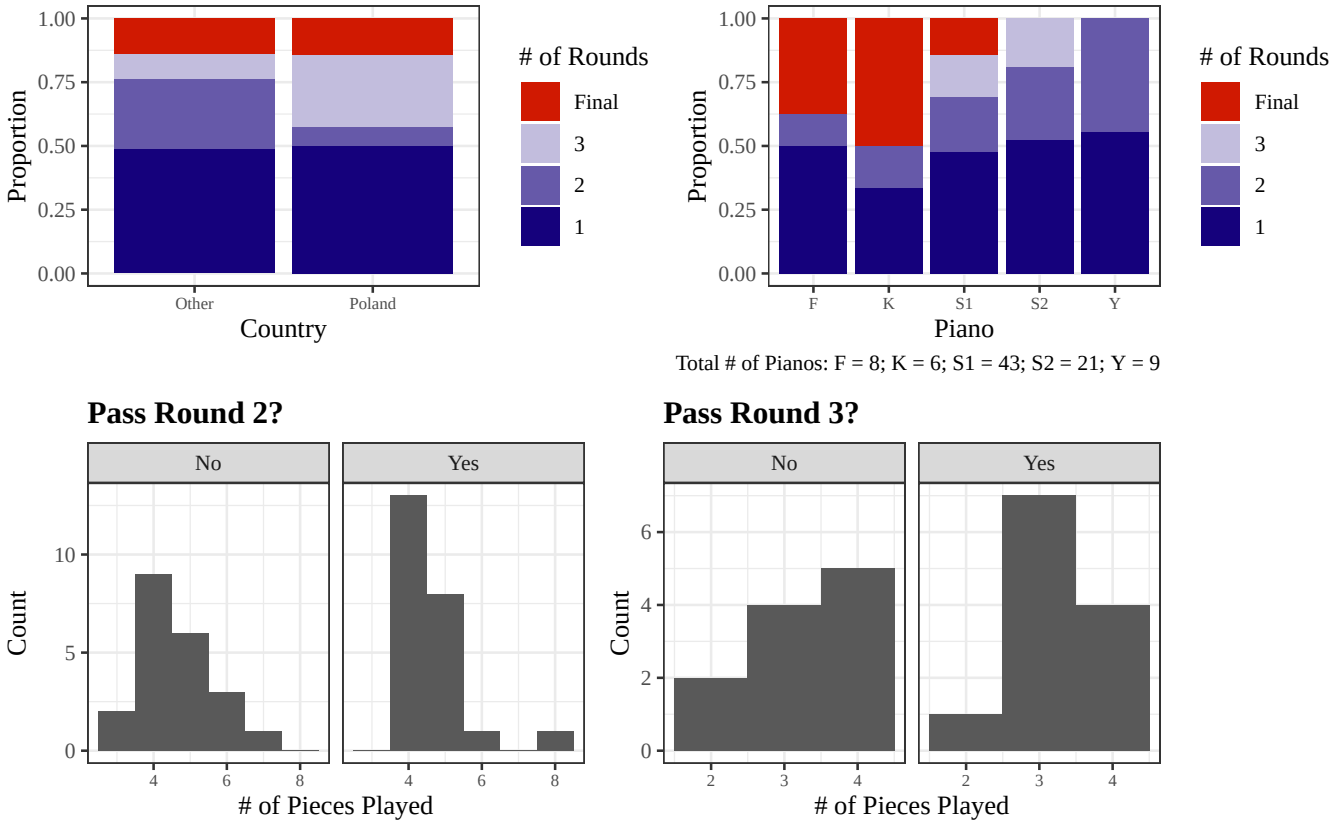
Above are the trace plots of the intercept and predictor variables in our model. For each variable, the random scatter around each respective mean coefficient value suggests that the chains have mixed well and that our model has converged.

A.4: Effective Sample Size and $\hat{R}$

Variable	Effective Sample Size	R_hat
(Intercept)	3556	1.000
Etude 1st Choice	4887	1.000
Barcarolle 2nd Choice	4782	0.999
Fantaisie 2nd Choice	4313	1.000
Scherzo 2nd Choice	4732	0.999
Day 2	4034	1.000
Day 3	3773	1.000
Day 4	3761	1.000
Day 5	3714	1.001
Morning Session	4967	1.000
Sigma	4310	1.000

Above is a table of effective sample sizes and $\hat{R}$ values for each of our coefficients. We have 5000 total MCMC posterior samples. Since the effective sample sizes are all fairly large and close to our true posterior sample size of 5000, our MCMC samples appear to be weakly correlated. Furthermore, all of the $\hat{R}$ are extremely close to 1 which suggests that our model has converged well.

Appendix B: Initial Analysis of Potential "Repertoire Effects," consisting of home-field advantage, piano type, and the number of pieces played per round



We wanted to explore if "home-field advantage" could be an additional effect to consider, given the competition is held in Poland and that could lead to biases from judges. As shown in Figure 1, a larger proportion of competitors reached the 3rd Round than the rest of the field; however, the proportions of those eliminated after Round 1 and those that made the finals are very similar. The right panel makes a similar comparison of Piano type to illustrate the proportion of competitors that chose each piano type that got eliminated in each round. As you can see, 50% of the competitors who chose the Kawai piano made it to the final round, and the lowest proportion of those competitors got eliminated in Round 1. On the other hand, none of the competitors that chose Yamahas or S2 (a type of Steinway) advanced to the final round. Due to the differences in the proportions in Piano type, this may be a variable we dive deeper into in our analysis.

We also compared the number of pieces played, as in the second and third main competition rounds, the number of pieces varied between competitors, and we hypothesized that could be associated with better or worse overall outcomes. For Round 3, for example, it seems that those who chose to perform three pieces had better outcomes than those who performed either two or four. As you could see, 11 competitors played three pieces and 7 of them advanced, while for those who chose to perform two or four pieces, more competitors were eliminated than advanced to the final round.

Appendix C: Variable Descriptions

Variable(s)	Explanation	Missing Values
First_Name and Last_Name	These variables list the first and last name, respectively of the contestant. In order to aid in dataset merging, we replaced special characters like accented vowels with their ASCII equivalents.	Since there are no missing values within this variable, no wrangling for null values was necessary.
Country	This variable lists the country that the contestant represents. Contestants came from all over the world, with a majority coming from North America, Europe, and Asia.	Since there are no missing values within this variable, no wrangling for null values was necessary.
Piano	This variable lists the piano model used by each contestant. Participants could use either a Steinway & Sons, Yamaha, Fazioli, or Kawai piano, with there being two Steinway options.	Since there are no missing values within this variable, no wrangling for null values was necessary.
P_x	These variables list the opus (and number if available) of the piece played by each contestant in the Preliminary. It is ordered such that P_1 refers to the first piece played, and P_6 refers to the last piece played (only 6 pieces were played in the Preliminary round). If a contestant skipped the Preliminary round, each piece is listed as "bye".	Since there are no missing values within this variable, no wrangling for null values was necessary.
P_x_piece	These variables list the genre of each piece in the Preliminary, with P_1_piece referring to the genre of the first piece played by each contestant and so on for a total of 6 pieces.	Contestants who joined the competition in Stage I have null values for each of these variables.
Pass_P	This variable is either "Yes" or "No" and refers to whether the contestant passed the Preliminary round. All values are "Yes", including for participants who did not perform during the Preliminaries.	Since there are no missing values within this variable, no wrangling for null values was necessary.
P_day	This variable lists the day the performance occurred, with 1 referring to the 1st day of the Preliminary round, and 12 referring to the last day.	Contestants who joined the competition in Stage I have null values for this variable.
P_time	This variable lists whether the performance occurred during the morning session or the evening session of the Preliminary round.	Contestants who joined the competition in Stage I have null values for this variable.
S1_x	These variables list the opus (and number if available) of the piece played by each contestant in Stage I. It is ordered such that S1_1 refers to the first piece played, and S1_4 refers to the last piece played (only 4 pieces were played in Stage I).	Since there are no missing values within this variable, no wrangling for null values was necessary.
S1_x_piece	These variables list the genre of each piece in Stage I, with S1_1_piece referring to the genre of the first piece played by each contestant and so on for a total of 4 pieces.	Since there are no missing values within this variable, no wrangling for null values was necessary.
S1_choice3 and S1_choice4	These variables refer to Stage I category choices. Since contestants had to choose an etude from categories 1 and 2, but could choose either a nocturne or etude from category 3 and either a ballade, barcarolle, fantaisie, or scherzo from catgeory 4, this variable was created to list which of the two the contestant chose from category 3 and which of the 4 the contestant chose from category 4, respectively.	Since there are no missing values within this variable, no wrangling for null values was necessary.

Variable(s)	Explanation	Missing Values
S2_x	These variables list the opus (and number if available) of the piece played by each contestant in Stage II. It is ordered such that S2_1 refers to the first piece played, and S2_8 refers to the last piece played (up to 8 pieces where played in Stage II, with a majority of contestants playing either 3 or 4 pieces).	Most of the missing values constitute people who did not make it to Stage II, with the remaining missing values consisting of people who did not play piece 4, 5, 6, 7, and 8.
S2_x_piece	These variables list the genre of each piece in Stage II, with S2_1_piece referring to the genre of the first piece played by each contestant and so on for a total of up to 8 pieces.	Most of the missing values constitute people who did not make it to Stage II, with much of the remaining missing values consisting of people who did not play piece 4, 5, 6, 7, and 8. Since contestants were allowed to play any other pieces by Chopin to meet the required performing time of 30 to 40 minutes but there is no way to confirm genres for those opuses, certain pieces do not have a genre [@zotero-101].
S3_x	These variables list the opus (and number if available) of the piece played by each contestant in Stage III. It is ordered such that S3_1 refers to the first piece played, and S3_4 refers to the last piece played (up to 4 pieces where played in Stage III, with a majority of contestants playing either 3 or 4 pieces).	Most of the missing values constitute people who did not make it to Stage III, with the remaining missing values consisting of people who did not play piece 3 and 4.
S3_x_piece	These variables list the genre of each piece in Stage III, with S3_1_piece referring to the genre of the first piece played by each contestant and so on for a total of up to 4 pieces.	Most of the missing values constitute people who did not make it to Stage III, with much of the remaining missing values consisting of people who did not play piece 3 and 4. Since contestants were allowed to play any other pieces by Chopin to meet the required performing time of 45 to 55 minutes but there is no way to confirm genres for those opuses, certain pieces do not have a genre [@zotero-101].

Variable(s)	Explanation	Missing Values
Yes_Sx	These variables refer to the proportion of "Yes" verdicts received from the jury, with Yes_S1 referring to the proportion received in Stage I, and Yes_S3 referring to the proportion received in Stage III for a total of 3 stages.	Here, missing values constitute people who did not make it to Stage II or Stage III, with more missing values as contestants get weeded out.
Pts_Sx	These variables refer to the adjusted average number of points received from the jury, with Pts_S1 referring to points received in Stage I, and Pts_S3 referring to points received in Stage III for a total of 3 stages. When calculating the average score, "scores were bounded within 3 points (Stage I) or 2 points (Stage II and III) of the true average; for example, if the true average were 14.35 in Stage I, all scores lower than 12 would be adjusted to 11.35 and all scores higher than 17 would be adjusted to 17.35, and the average would then be calculated again".	Missing values constitute people who did not make it to Stage II or Stage III, with more missing values as contestants get weeded out.
Pass_Sx	These variables are either "Yes" or "No" and refers to whether the contestant passed each stage, with Pass_S1 referring to whether the participant passed Stage I, and Pass_S3 referring to whether the participant passed Stage III for a total of 3 stages.	Missing values constitute people who did not make it to Stage II or Stage III, with more missing values as contestants get weeded out.
Sx_day	These variables list the day the performance occurred, with S1_day referring to Stage I (and so on until Stage III) and a 1 referring to the 1st day of the each of the stages.	Missing values constitute people who did not make it to Stage II or Stage III, with more missing values as contestants get weeded out.
Sx_day	These variables list whether the performance occurred during the morning session or the evening session of each of the stages, with S1_day referring to Stage I and so on.	Missing values constitute people who did not make it to Stage II or Stage III, with more missing values as contestants get weeded out.
F_1	This variable lists the opus of the concerto played by each contestant in the Final.	Missing values constitute people who did not make it to the Final.
Final	This variable refers to the final rank of the competitors, with numbers 1-6 referring to contestants who placed (from 1st place to 6th), and "F"" referring to contestants who did not place.	Missing values constitute people who did not make it to the Final.
F_day	This variable lists the day the performance occurred, with 1 referring to the 1st day of the Final round, and 3 referring to the last day.	Missing values constitute people who did not make it to the Final.

Response to Peer Review

Introduction

The suggestion to reorder the graphs in figure one is a good one. Our original intention was to group each of the figures by session and day (which then makes the prior ordering make sense), but when placing the two groups together, the grouping is lost. As such, the order suggested makes more sense and helps the visualization flow better. Furthermore, we understand the hesitance in using "loser" to describe a group people. We were hoping that the catchyness would make up for the harshness of the word, but the term is a bit too crass for a more formal setting. As such, we changed "loser" to "unlucky." We also replaced raw variables names and went less indepth in our description of relevant predictors. Not only is such detail unnecessary since we have a description of all variables in the Appendix, a broad overview leaves more space to describe more critical aspects of our report.

Methodology

The use of ϵ and the lack of a hat over our response variable is an oversight on our end. While the purpose of our analysis is not to predict adjusted average points in Stage I, we find that including the estimated beta coefficients aids in continuity between the methodology and results sections. The clarification regarding what each estimated β refers to is still necessary however, so we took your suggestion and included that the estimated β's refers to posterior means. We also clarified that our expected outcome is conditional (for those not familiar with the Bayesian framework). Furthermore, the questions you raised in the last paragraph of this section are all very valid questions in terms of reproducibility, so we included answers in the body of our work instead of just directing readers to the Appendix.

Results

We removed the extraneous results, such as the log-posterior, in order to focus more clearly on the posterior mean estimates and credible intervals of our final model. Regarding the Bayesian interpretation of our results, we went back through our explanations of the posterior mean estimates and added discussion of the probabilistic nature of their 95% credible intervals.

Discussion

Similar to the results section, we took the suggestion to better explain the Bayesian interpretability of our results into consideration by adding a probabilistic explanation through the use of the framing of the credible intervals. This also allowed us to better explain our thoughts regarding the potential significance of the factors we were considering in the final model. Additionally, we explained in greater detail which specific variables were not included due to poor mixing and the potential effects this decision would have. We also added a final "summary" section, as suggested to summarize bottom-line results and important conclusions, as well as connect the study back to the overall goals of the project.