Mutilevel Analysis on the Battleship Numberline Study Part I

Zuojun Gong

November 18, 2016

EDA

There are total of 20 questions of numbers between zero and one, and there are total of 8257 attempts on the questions. We have data from 414 participants in our experiment and they were involved in 15 types of experiments that each contains one or several types of questions. The subjects were given at most 10 seconds the respond to the questions and their accuracy of their solution is judged and they were given the feedback. Additionally, the subjects are able to keep track of their total average accuracy and the number of questions answered right as well. In this experiment, we also track data including whether the guidance is enable, their response time, how well they answered the questions, and the decimal value of their guesses.

Nearly a third of the players attempted all 20 questions, and the rest of the players tried a relatively smaller amount of questions, with more than a third ranging between one to five questions. We can see that most of the players had only one experiment, and only 7% of the players had more than one type of experiment during the study. All of the experiments included all types of questions. When looking at the relationship between players and their IP address, our results suggest that 373 players used only one IP address to play this game, whereas 383 players have the same partial IP addresses. This suggest that some of the players played the game while using different IP addresses, and therefore we want to decide whether we want to take that into factor and control for the IP address when we model our data. Especially considering we don't know if the change of IP addresses has an impact on player experience and performance.

When we examine the distribution of proportion-correct scores for all the players, we can see that most of the players have a low accuracy in making a correct answer - a perfect hit, since histogram appears to be right skewed. In terms of fraction of the answers that are correct for each of the questions, we can see that it appears to decrease as it approaches to the middle, with the exception of 0.5. Upon examining if guide played a role in the improved accuracy, we determined that it is very unlikely that the guide marks was a factor in high accuracy at 0.5 since the histogram appears to be similar between the total accuracy and the unguided accuracy.

When examining the reaction time across all players, we first remove the case where the reaction time was 46.6 seconds, as it is a leverage point for our graphical analysis. Given the subset of the data, we can see that across the players there is a lot of variation in the reaction time across players, but the variation in the reaction time across questions appears to be very little.

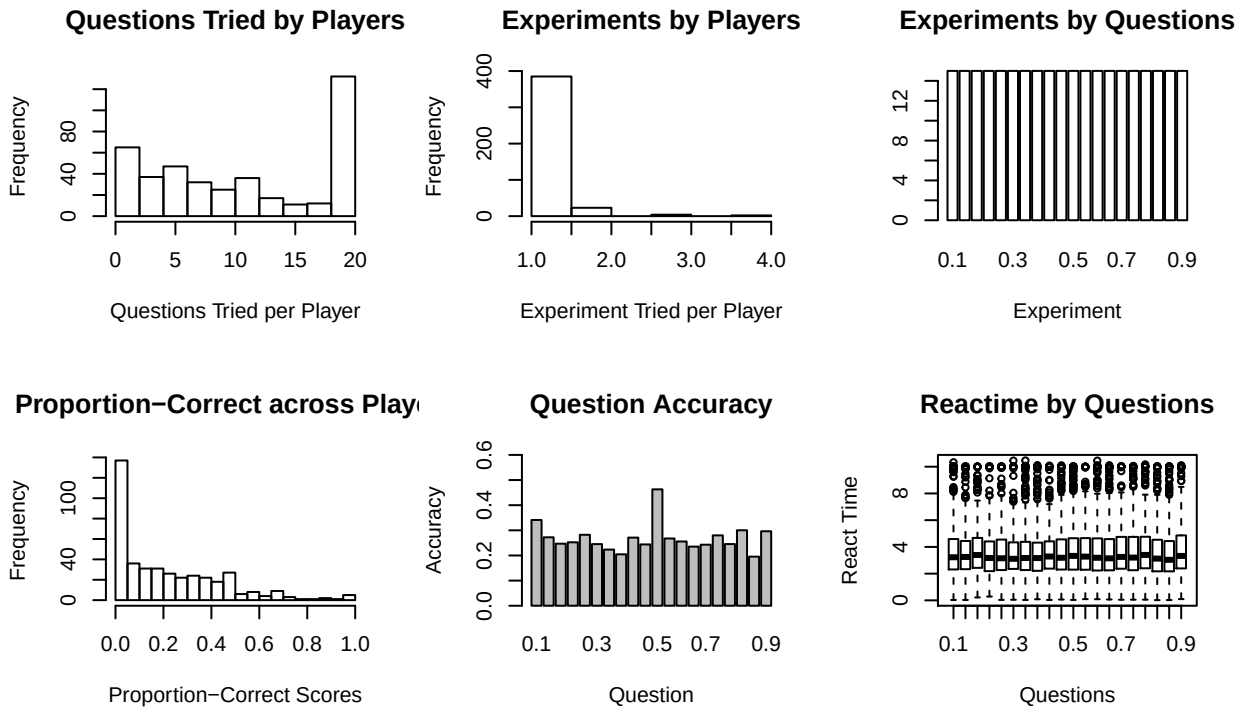
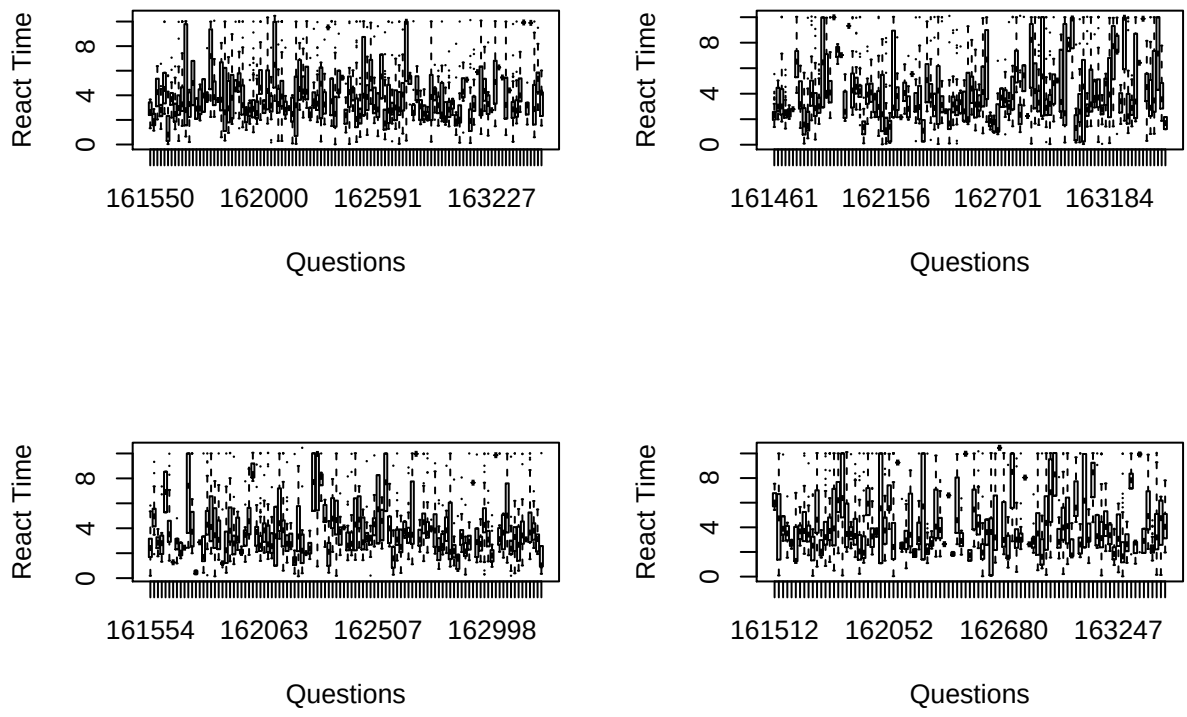


Figure 1: Univariate Analysis between the relationship of Questions and Players, Experiments and Players, Experiment and Questions, Porportion Correctness across Players, Question Accuracy and Reaction Times



Logistic regression for question difficulty

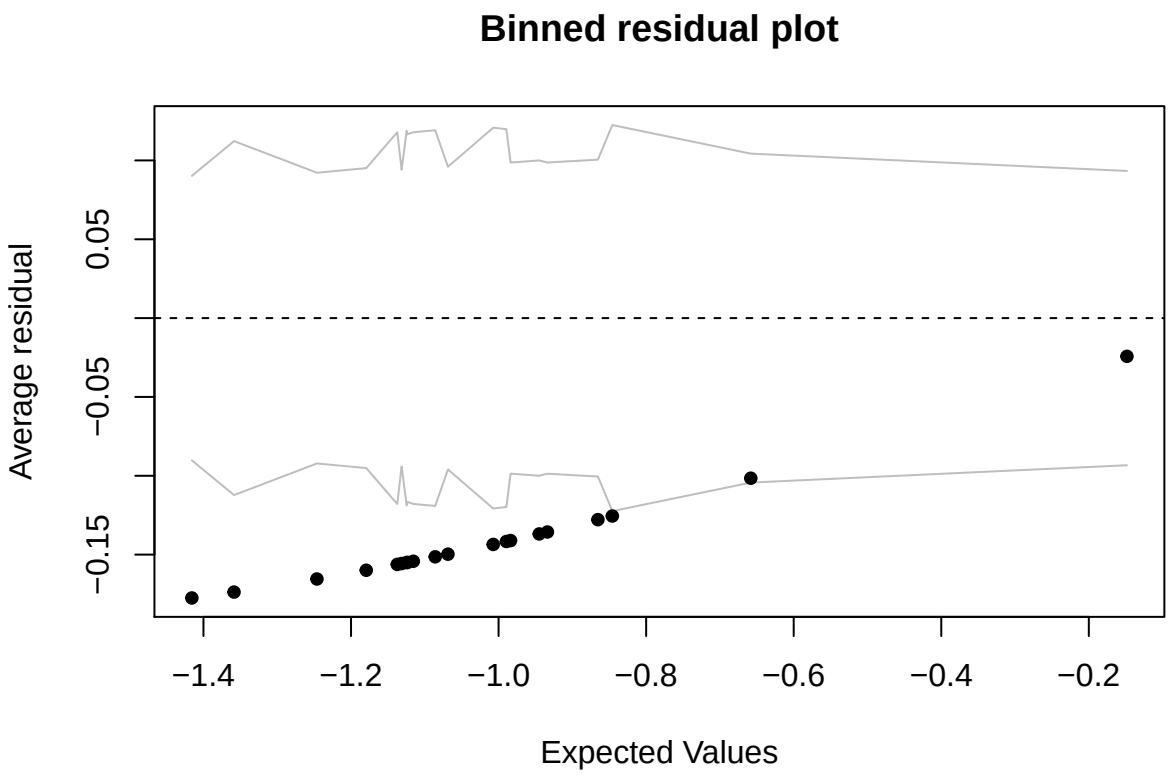


Figure 2: Binned Residual Plot for Logistic Regression on Question Diffculity

First, we fit a logistic regression model giving the probability that a question will be answered correctly, using the difficulty of the question as a factor. Our binned residual plot shows poor fit for our data, as most of the residuals display a patter and is not close to the zero line. The parameters of our model is the inverse logit of the probability that a given question is answered correctly. For example, our model predicts that the probability of answering question 0.1 right is 0.3411. By taking the intercept out from the model, we make it easier to compare the coefficients across all current questions, which is a factor variable. This allow us to compare the probability of any given factor to the null: P(correct) = 0.5 rather comparing to the P(correct|currentQuestion = 1). Therefore it is much easier for interpretation.

We tried both using the raw fraction of the participants' accuracy and the logit of the accuracy, and determined that the logit provide us with a better plot. We argue that the coefficient of estimates of our logistic model directly translate to the predicted probability of participant correctness on a particular question, in an inverse logit manner. In other words, the predicted probability of correctness of our model is the inverse logit of the coefficient of estimate. Since the raw fraction of participant correctness is an estimator of the probability of correctness, the logit of the fraction is comparable to our coefficient of estimate.

Our plot suggests that the predicted probability from our model is the same as the fraction of players who got the question right. This means that our model is using the average of correctness for each question as estimator, and our estimation is unbiased.

We then attempt to improve our model by variable selection. We first selected several variables that from the datasets that can help contribute extra information to our model. In addition to the type of question, we also considered the whether guide is enabled, the current level number, the average accuracy of the player at

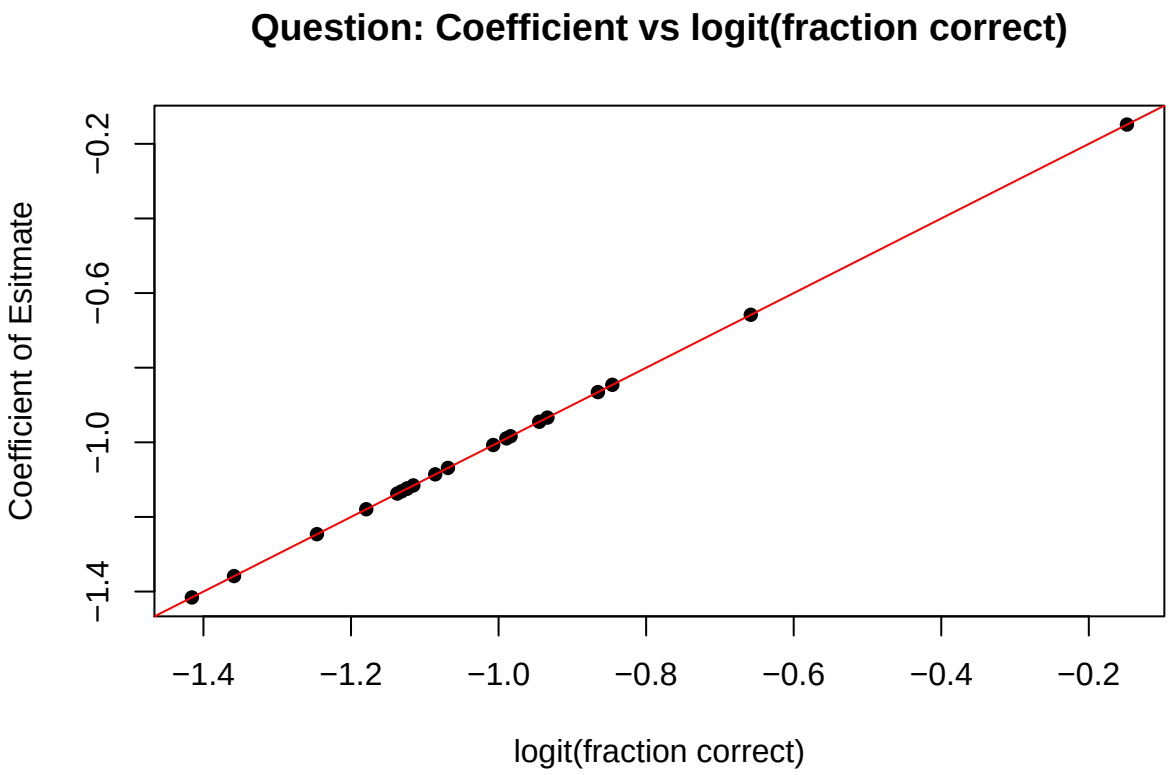
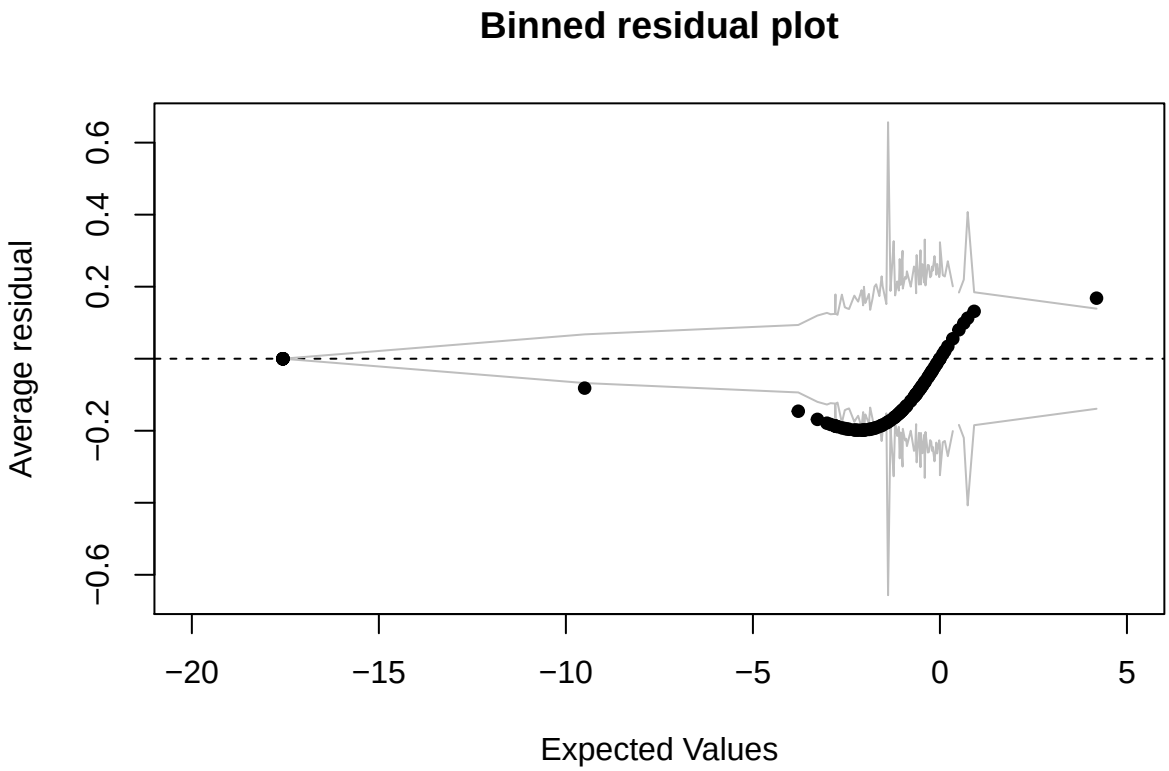


Figure 3: Coefficients against the fraction of participants who got the corresponding question right

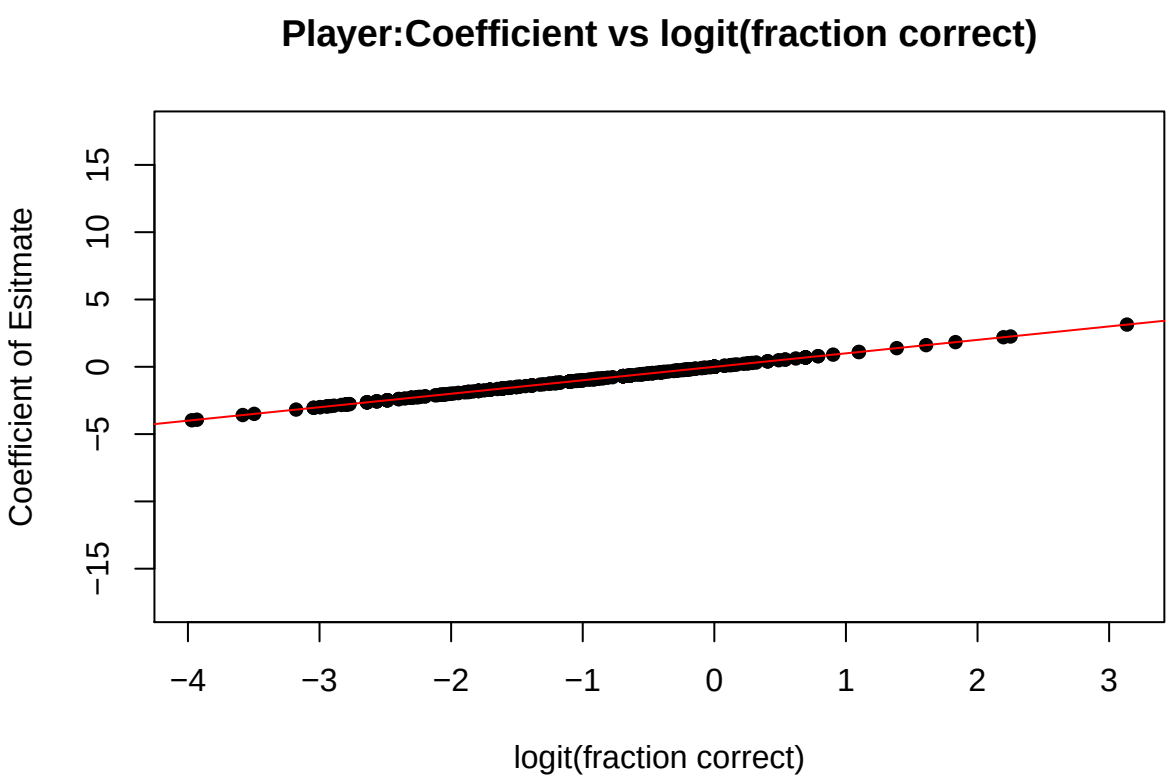
the time, and the reaction time of the player. Then we used stepwise AIC to search for a possible model that minimizes the AIC. After the procedure, we find out that dropping current level number will improve our model, and the chi-squared test suggest that there is no evidence against the smaller model, thus we will go with the reduced model.

We interpret our parameters for the logistic regression by using the divide by four rule. We can interpret the intercept as: Given a player answers question one without guide enable, with zero percent average accuracy, and zero seconds of reaction time, he has a probability of invlogit(-7.70516) = 0.00045 getting the answer right. The coefficient Question 0.13 divided by four means that holding all other variable constant, the probability of getting question 0.13 right is 0.11 lower than the probability of getting question 0.10 right. This can be applied to all the factor coefficients for questions in our model. If the guide is enabled, a player has 0.107 high probability of having the correct answer, holding all other variables constant. And for each additional average accuracy percentage, it is correlated with a 0.02 increase in probability of answering correctly, holding all other variables constant. And for each additional reaction time in seconds, the probability of answering correctly increase by 0.0074, holding all other variables constant. This model suggests that individual performance is an important factor in answer accuracy.

Logistic regression for player proficiency



Then, we want to examine the relationship between the the probability that a player will provide a correct answer and player proficiency by fitting a logistic regression model. Our binned plot shows a better fit of this model compared to the previous one. This is consistent with our hypothesis during the EDA: there was very little variation in the average fraction of a question being answered right, but there was more variation in the average fraction of a player's correctness. The coefficient of estimate is the inverse logic of the predicted probability of a player's accuracy. For instance, player 161461 is estimated to have a probability of 0.6874 answering a question right.



Similar to the previous model with current question as a factor, we tried both using the raw fraction of the proportion correct for each player and the logit of the accuracy, and determined that the logit provide us with a better plot. We argue that the coefficient of estimates of our logistic model directly translate to the predicted probability of participant answers correctly on any question, in an inverse logit manner. In other words, the predicted probability of correctness of our model is the inverse logit of the coefficient of estimate. Since the raw proportion correct for each player is an estimator of the probability of correctness for individuals, the logit of the fraction is comparable to our model's coefficient of estimate.

Our plot suggests that the predicted probability from our model is the same as the fraction of players who got the question right. This means that our model is using the average of correctness for each question as estimator, and our estimation is unbiased. Note that our plot does not include the average fraction correct for those players who never got a single question right, therefore they are not represented in this plot.

Mixed Effects Models

First, we fit a mixed effects logistic regression predicting the probability of a correct response, using question diffculity as a fixed effect factor, omitting the intercept, and with a random intercept grouped by player ID.

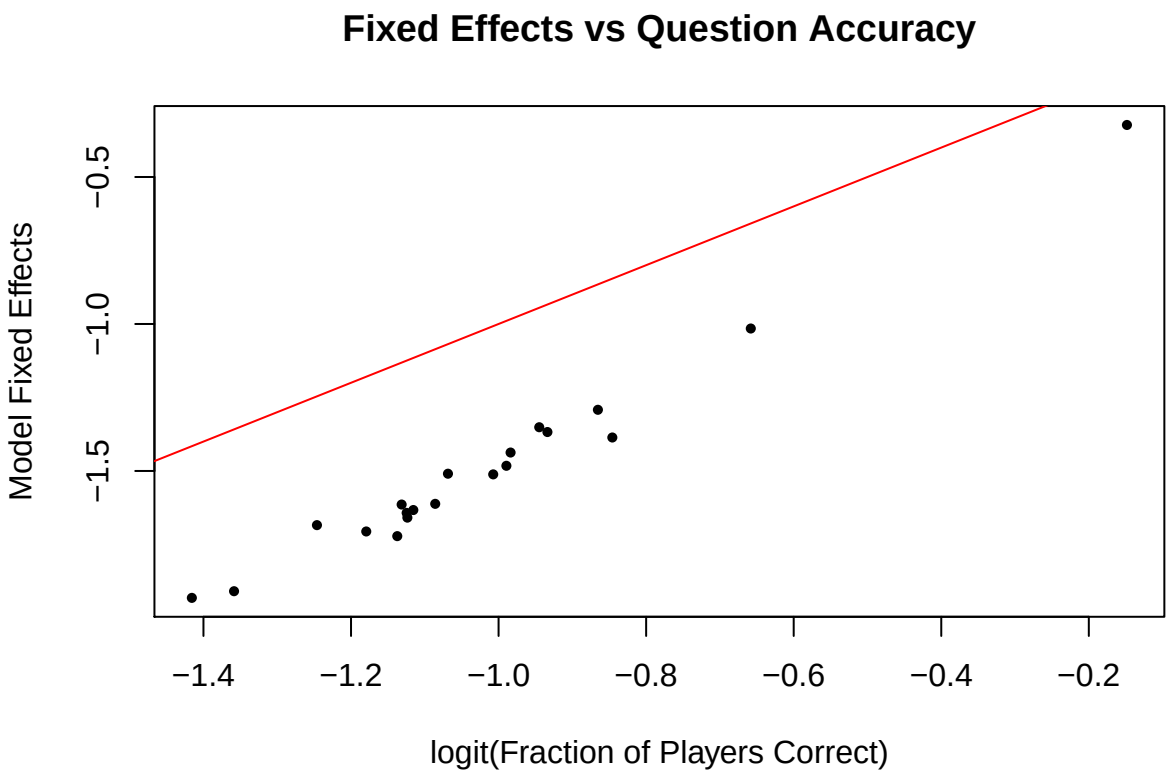


Figure 4: fixed effects from this model against the fraction of players who got the corresponding question correct

The fixed effect of our model represents the effect of questions on probabilities of answering correctly across all individuals, and from our plot we can see that coefficient of the fixed effect of our model is smaller than logit(Fraction of Players Correct on Question). Our model suggests that the effect of questions on the probability of player answering correctly is lower than the fraction of players that answers correctly. This means that the effect of the questions is smaller than what it appears from the unbiased estimator, and that individual effects play an important role in the answer accuracy.

The random effect of our model represents the effect of individuals on the probability of answering accuracy. In the mix model we partially polled the individuals' effects, and the plot suggest that the individual's effects are more significant in determining how question answer accuracy than the questions themselves. For instance, the for participant one, the probability of this person answering each question correct respectively can be calculated by adding the random effect to fixed effect, and take the inverse logit.

We then try to improve our model by replacing or augmenting currentQuestion with other variables, judging by the AIC, BIC and DIC and the residual plots.

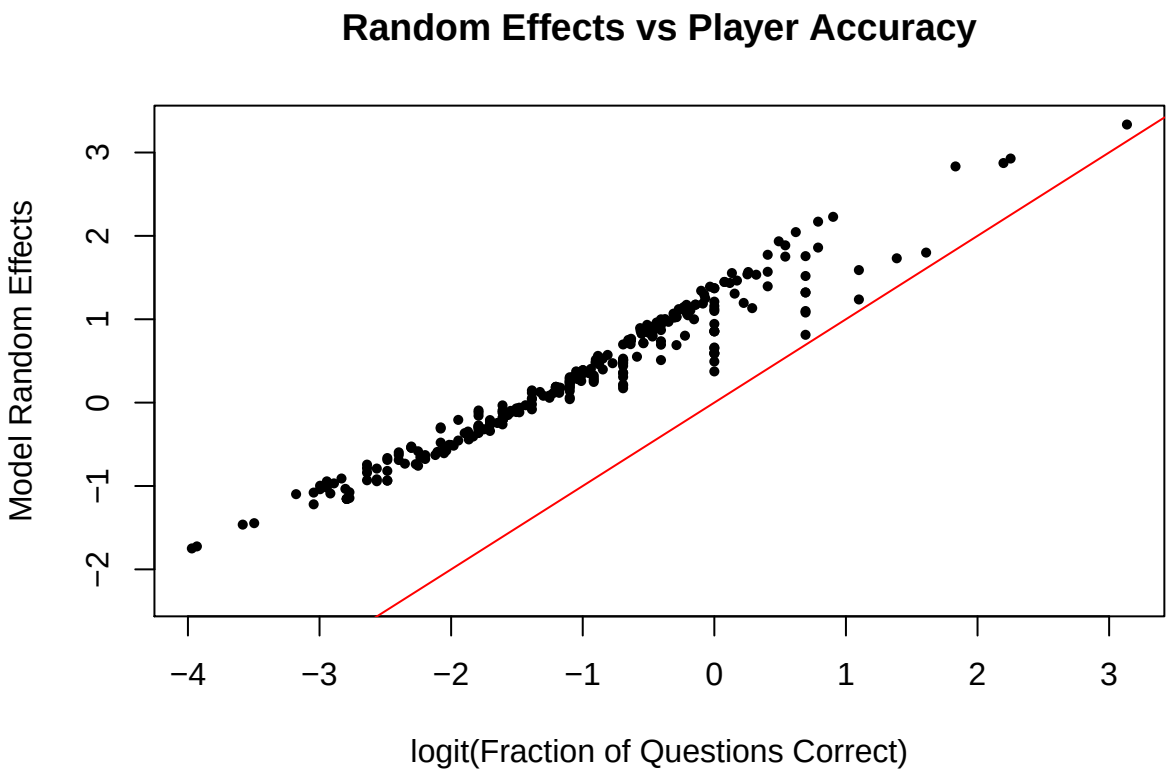


Figure 5: random effects against the proportion correct for each player

Table 1: AIC, BIC, DIC Compairson for the Six models

	M1	M2	M3	M4	M5	M6
AIC	8083.054	8083.054	8477.553	8613.136	7886.789	7890.757
BIC	8118.148	8118.148	8519.666	8760.531	8055.241	8087.284
DIC	7031.529	6978.379	6991.847	7275.638	7275.637	6379.870

First of all, we only consider models that converges and have reasonable eigenvalue ratio. Then from the valid models, We used model AIC, BIC and DIC to with selected variables to determine which model is the best option. We have six models, and eventually we ruled out the models that includes current questions, as they do not converge. We also scaled average accuracy and reaction time to improve eigenvalue ratio, and eventually we are picking between two models. M5 uses participant ID as varying intercept and use whether Guides is Enabled, scaled average accuracy and scaled reaction time as fixed effects. M6 uses participant ID as varying intercept and scaled average accuracy as varying slope, and whether Guides is Enabled and scaled reaction time as fixed effects. It appears that M5 is better in AIC and BIC, however M6 is better in DIC. Upon examining the residual plots, we determine that M5 is a better fit for our data.

Given M5, we want to see whether adding a second random intercept corresponding to ip, ip2, or ip3 helps.

Table 2: AIC, BIC, DIC Compairson for Three models with IP and SID, and Three without SID

	M5.ip	M5.ip2	M5.ip3	M5.ip.a	M5.ip2.b	M5.ip3.c
AIC	8094.677	8104.904	8088.294	8075.082	8076.083	8073.067
BIC	8129.771	8139.998	8123.388	8117.195	8118.196	8115.179
DIC	7268.269	7284.897	7293.387	7488.794	7602.430	7503.898

Adding the IP address does not improve our model. And by having both the IP address and the SID in our model, we have redundent information, as we discovered in EDA that most IP addresses are unique. We can see that by removing the SID from the model, all models including IP addresses improve. However, this improvement is not sufficient to show that the model is better than the previous model that does not include IP address, therefore we decide to continune with our model M5.

Summary

We used the data collected from the battleship numberline game that contains 8257 trails from 414 participants. We included information such as player ID, name of the experiment, question type, player reaction time, whether guide is enabled, name of the level, whether the response is correct, average reaction time, average precision and etc. We want to see if we can use covariates to predict the probability of a player giving the correct answer on a given question. Particularity, we want to see if the difficulty of the question has an impact on the probability of the player providing the right answer.

Our exploratory analysis provided us some evidence of correlation between the variables. We observed that the fraction of players that answered the question correctly varies by the question slightly, and the highest accuracy occurs when the number is 0.5. We also observed that player's accuracy on all questions is distributed with a right skew. Most of the participants had an accuracy closer to zero. We also investigated the reaction times across players and across questions, and found out that individual players vary greatly in reaction times but overall reaction time on questions are similar.

Our initial logistic models investigated the relationship between the difficulty of the question and the probability of the question being answered correctly, and the probability of a player answering any question correctly, considering information from other variables. Our models provide us with an insight that the question difficulty, current level number, average accuracy and reaction time in effect in predicting the probability of a player answering the question correctly. However, the diagnostics suggest that this model has a poor fit on the data. Our logistic regression with player ID as a covariate fits the data better. We speculate that the difficulty of the questions does not play as important of a role in helping us to predict the probability correct answers. Additionally, we are interested to see how the individual proficiencies of the participants affect the probability of giving correct answers.

We want to use a partial pooling model to estimate the effect, and we achieve that by fitting a mixture model on our data set with random effects (i.e. varying intercepts). Our initial mixture model predicts the probability of answering correctly, with question difficulty as a fixed effect factor, and the individual player as random effects. This model uses question difficulty as a fixed information for all players, but it also takes in consideration from how the individual players preform when answering the questions. Our model suggests that there is a bigger influence on the probability of correct answers from the player's individual performance than from the difficulty of the question itself. We then continued to search for a better mixture model by minimizing the AIC, BIC and DIC of the model as well as finding a model that converges. We also used scaling on continuous variables to improve our model fitness. We discussed the possibility of using IP address instead of player ID, but decided against it due to its lack of model improvement.

After model selection, we determined that our final model is a mixture model that uses player ID as random effects, and uses whether the guide is enabled, the scaled average accuracy of the player, and the scaled reaction time of the player as fixed effects. Our model's fixed effect parameters suggest that on average, enabling the guide is associated with a 7.8 percent increase in probability of a player getting the correct answer. Every percentage increase in player average accuracy is associated with an additional 1.77 percent increase in probability of a player getting the correct answer. And for every second increase in player reaction time, the average probability of a player getting the correct answer increases by 0.12 percent. Our model concludes no significant relationship between the predicted probability and the difficulty of the question, and thus we cannot conclude that there is a relationship between the probability of a player answering the question correctly and the difficulty of the question.

Appendix I

Gelman, A. & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. NY: Cambridge Univ Press.

Lomas, D., Patel, K., Forlizzi, J.L., and Koedinger, K.R. (2013). Optimizing Challenge in an Educational Game Using Large-Scale Design Experiments. Paper presented at CHI 2013, Paris, France. Obtained online at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.480.2493&rep=rep1&type=pdf

Lynch, Scott M. (2007). Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. New York: Springer.