Strategy in StarCraft 2 Learning Curves

by

Justin O’Camb

B.A., University of Nevada, Reno, 2019

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Psychology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY
Summer 2022

Declaration of Committee

Name:	**Justin O'Camb**
Degree:	**Master of Arts**
Title:	**Strategy in StarCraft 2 Learning Curves**
Committee:	**Chair: Ralph Mistlberger** Professor, Psychology
	Mark Blair Supervisor Associate Professor, Psychology
	Matthew Sigal Committee Member Lecturer, Psychology
	Joseph Thompson Committee Member Instructor, Psychology
	Parmit Chilana Examiner Associate Professor, Computing Science

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Abstract

The power law of practice has been a long-standing theory of the learning curve: how skill improves with experience. Despite the general agreement that learning fits a smooth curved pattern, there have been specific areas (piecewise learning curves, improper aggregation, and plateaus) that have refuted the existence of single smooth learning curves for a given task. The present study attempts to generalize strategy-specific learning curves to a large longitudinal dataset of games in StarCraft 2, a highly complex task with refined measures of skill. Using novel methodology to balance error across high dimensional measures, the ubiquity of the power law of practice is not supported, yet shifts in strategy do not account for the lack of power law learning curves. The existence of hobbyists (people who do not want to become experts) is considered as a possible explanation. Implications of the findings and directions for further research are discussed.

Keywords: Learning Curves; Power Law; StarCraft 2; Strategy; Hobbyist

Acknowledgements

I would like to sincerely thank Mark Blair, Matt Sigal, Joe Thompson, and all the members of the Cognitive science lab for help brainstorming and working through the many difficulties that arose while working with a complicated dataset. All the support and guidance I have received while working through this project has led me to more fully understand the field of psychology and research as a whole. I would also like to thank the friends I have made in the psychology department, specifically Robin Barrett, Brittany Dennett, McKenzie Braley, and Jourdan Jackson who all tolerated my incessant ramblings that ultimately helped shape the project into what it is now.

Table of Contents

List of Figures

Chapter 1.

Introduction

Skill in a given task is thought to increase along a curve over time with practice and experience (Anderson, 1982; Ericsson, Krampe, & Tesch-Römer, 1993; Keller, 1958; Newell & Rosenbloom, 1981). Newell and Rosenbloom (1981) suggest the power law of practice to fit general learning data. This theory states that performance measures (like reaction times) will fit a log-log-linear trend: meaning the log values of the performance measure will have a linear relationship with the log values of the trial number. This trend suggests that the untransformed data fits a power curve. Newell and Rosenbloom write extensively about the ubiquity and universality of the power curve in learning data. Although the power law of practice has been widely accepted and used as a sort of golden standard for learning data (Anderson, 1982; Ericsson, Krampe, & Tesch-Römer, 1993; Logan, 1988), the suggestion of a single smooth curve to model a person's experience in a task could be misleading. There is debate in the literature as to whether a single smooth curve is the best fit for a learning process. Many of these concerns can be grouped into three categories: piecewise curves (Delaney et al., 1998; Donner & Hardy, 2015; Rickard, 1997), improper aggregation (Gallistel, Fairhurst, & Balsam, 2004; Haider & Frensch, 2002; Myung, Kim, & Pitt, 2000), and plateaus (Bryan & Harter, 1897; Bryan & Harter, 1899; Gray & Lindstedt, 2016; Rahman & Gray, 2020).

1.1. Inconsistencies in Learning Curves

1.1.1. Piecewise Learning Curves

Piecewise learning curves refer to the separation of learning curves based on specific strategies or conceptualizations of the task. Rickard (1997) referred to claims in automaticity research (Compton & Logan, 1991) in which a shift in strategy is responsible for a task becoming automatic. Rickard fit separate power law curves to the different strategies from participants. The separate curves did not provide any significant reduction in R^2, despite Rickard claiming they provided a better visual fit. However, other research (Delaney et al., 1998) found that there was a much better fit when using piecewise power functions dependent on similar strategy changes as Rickard's study.

Delaney et al. (1998) showed that in previous simple arithmetic tasks, changes in strategy lead to the formation of new learning curves. Strategy-dependent learning curves led to a large increase in R^2 using the log-log-linear method of fit from the power law of practice. Strategies were defined as the change between mental calculation and memory retrieval, which represents a cognitive shift in the understanding of the task. To further test the influence of strategy on learning, Delaney et al. also re-analyzed a case study in which the participant was asked to change strategies at certain points throughout the experiment. Separate learning curves for each strategy led to a significantly better fit than an individual curve. The case study was also a simple arithmetic task, yet strategies were defined as the mental approach to solving the problem (e.g., starting with the largest digits, then working down). Although the analysis was just a case study, it shows this more general definition of strategy could also have separate underlying learning curves.

One drawback of the research in the area of piecewise curves is the simplicity of the learning tasks. The researchers have relatively clear and simple definitions for strategy which lead to clear separations of learning curves. Games on the website Lumosity offer a slightly more complex task than the other piecewise curve research. Analysis of Lumosity data has also shown that piecewise functions fit individual learning trends much better than a single power law curve (Donner & Hardy, 2015). However, the number of curves were picked based on what yielded the best fit. This exploratory method might suggest some underlying process causing separate curves but cannot necessarily be attributed to strategy for this study. Despite the better fits of strategy-dependent learning curves for simple learning tasks, piecewise functions are scarcely applied in learning studies. Depending on the aggregation methods used, there could still be a smooth visual fit for a single curve which would dissuade researchers from looking for any strategy related effects.

1.1.2. Improper Aggregation

Improper aggregation of learning outcomes can result in a smooth curve that may not accurately reflect the development of skill over time. Averaging across participants causes any dips and leaps in learning curves to be smoothed out in the final model. Because of this, individual participant learning curves should also be considered (Gallistel, Fairhurst, & Balsam, 2004; Haider & Frensch, 2002; Myung, Kim, & Pitt, 2000). Haider and Frensch also showed that changes in participant strategy could lead to shifts in performance that might not be visible in a full curve. The researchers suggested that the power law curve could be popular due to its robustness to underlying abrupt shifts in performance. In a continuation of Delaney et al., (1998), Haider and Frensch (2002) conducted an analysis of an alphabet verification task. Strategy was defined in a more general way as any time a participant shifts to a more optimal approach for the task. The task also included a simple and slightly more complex condition, despite still being a simple task overall. Finding a more efficient strategy led to shifts in participants' learning curves. This analysis suggests the findings of the piecewise literature can be generalized to broad definitions of strategy and possibly more complex tasks; yet Haider and Frensch did not fit a piecewise curve function, they only measured deviations from the log-log-linear fit.

1.1.3. Plateaus

Plateaus refer to periods of minimal or no improvement which cause deviations from power curves. Plateaus have been observed in more complex tasks, most infamously with training telegraph operators (Bryan & Harter, 1897; Bryan & Harter, 1899). Some of the telegraph operators in these studies showed distinct long plateaus in speed for receiving messages, followed by further improvement. Bryan and Harter (1899) explain this phenomenon through a theory of hierarchical conceptualizations of language that is very similar to modern chunking. Bryan and Harter suggested these steps forwards in understanding were related to the observed plateaus in their first study (1897).

There was debate as to whether the plateaus were a real artifact of learning or due to insufficient measurement. Keller (1958) stated that Bryan and Harter's testing methods were inadequate, which led to false plateaus among their participants. Keller also cited various other motor learning studies in which no plateaus occurred. However, in all the experiments graphed by Keller, data were aggregated across participants (which would smooth out any plateaus observable in each participant) or the participants did not practice long enough to reach asymptotic performance (their performance appeared linear rather than a power curve). With these issues in mind, it is still possible that plateaus occur in complex tasks.

Gray and Lindstedt expanded the above research surrounding the natural inconsistencies in learning. In their review, Gray and Lindstedt (2016) describe complex tasks as systems of subtasks that require a person to periodically adjust their strategy in order to continue to improve (get past plateaus). After facing a plateau, the person would search for a new strategy that would succeed in more subtasks than the current strategy. While the person searches and tries various strategies, they face a temporary dip in skill. Then, when they settle on a new strategy, the person will leap in skill while they learn and improve with the new, more optimal strategy. This explanation is a continuation of Brian and Harter's suggestions where new conceptualizations by the participant allow for the adaptation of a new strategy which can end a learning plateau and result in a leap in skill. Strategy-based improvement in learning was then shown in practice with the computer game Space Fortress (Rahman & Gray, 2020); changes in participant flight paths explained dips and leaps in performance. The game is based off of an arcade game from the 1980s (Star Castle). Although the graphics have been modernized in some cases, the core mechanics have remained the same. The game involves a hexagon in the middle of the screen with a rotating turret on the inside to represent the fortress. There is a larger hexagon surrounding the fortress which represents the boundary of the flight area. The player flies a spaceship within the flight area and shoots at the fortress to slowly take down its health, meanwhile the player must dodge shots and mines that are fired from the fortress. Space Fortress offers a more complex task to further test plateaus and generalize strategy-based deviations from a smooth learning curve. Rahman and Gray showed that strategy further generalizes to a broader definition of the approach method used by the participant.

1.2. Considering Task Complexity

Piecewise curves, improper aggregation, and plateau research have all indicated that there are definite strategy-based shifts from the golden standard power law of practice. The clearest of these occurred in the piecewise curve research with simple tasks and simple definitions for strategy. When researchers generalized to more complex tasks and more broad definitions of strategy, the shifts were not measured as piecewise curves, but rather general shifts from the power curve. Measuring general shifts also required more complicated methods than just piecewise curves; yet piecewise curves have not been generalized to complex tasks in the same way. If strategy-based separate curves generalize to complex tasks, this would offer a much more methodologically simple way to model deviations from power curves.

Testing piecewise curves on a variety of tasks is important since there is evidence to suggest that simple learning tasks in a lab setting do not always generalize to complex tasks. Robertson and Glines (1985) showed concerns that simple motor tasks do not yield enough complexity to require a hierarchy of conceptualizations. This statement is in line with other research that suggests simple tasks in laboratory studies require low cognitive load and do not generalize when compared to learning complex motor tasks (Wulf & Shea, 2002).

In order to test the generalizability of strategy-dependent curves, a complex task with a clear definition of strategy is needed. There also needs to be enough data to capture the full learning curve of the participants, which can take a long period of time for complex, real world tasks. Video games offer a viable platform for this sort of testing. There are many difficult games that take a long time to master. With games played on a computer, all the commands can be monitored and used to find a proper definition for strategy. StarCraft 2 is a game that offers ample complexity, is played on a computer where inputs can be precisely measured, and players can download the logs of all games played for their account which allows for long-term longitudinal data collection. Therefore, StarCraft 2 is a strong option to test the generalizability of strategy-dependent learning curves.

1.3. StarCraft 2

StarCraft 2 is a highly complex game which could provide a platform for more in-depth analysis of strategy dependent piecewise power law functions. With the lack of a consistent enemy, players' development of multiple powerful strategies becomes vital. In StarCraft 2 games, the main goal is to destroy the enemy base. Players can pick from one of three races (Protoss, Terran, or Zerg). They are then matched against another player of similar skill level. Both players only see a limited portion of the map at any one time, but can use their mouse to move the screen to different areas. The player then uses keyboard shortcuts in order to command workers, create buildings, and build military units. They often send a few of their soldiers to the enemy base to scout out what buildings and military units the enemy player is building. From here on out, the game becomes a battle of speed and military strategy for either player to attempt to gain an advantage over the other. Figure 1.1 shows the start of a new game of StarCraft 2. In the center of the image is the starting base for Terran players. The minimap is located in the bottom left corner with the area outside of direct vision of any units grayed out. One of the worker units (SCV) is selected, so the unit is shown in the bottom middle of the screen, with the build menu on the bottom right. Each building has a keyboard button associated with it so player can give commands faster than individual clicking menu options. As the player expands, scouts their enemy's base, and places more buildings, they must move the camera to see different areas of the map.

Figure 1.1. Screenshot from StarCraft 2

In a cross-sectional StarCraft 2 study, Thompson, Blair, Chen, & Henrey (2013) identified first action latency (FAL) as a strong predictor of skill across all levels of beginner and expert play. FAL was defined as the time it takes a player to perform an action after moving the screen. This measure is calculated with every occurrence of a screen move, then averaged for each game. Because there are a large number of FALs per game, the averages have a reasonable consistency from game to game. This measure was thought to be heavily influenced by the speed, reaction time, and automaticity of players when carrying out certain sets of actions. Therefore, this measure will be important in analyzing the skill increases and decreases with any given strategy over time.

A longitudinal study later considered FAL as a primary skill measure. In his dissertation, Thompson (2018) focused on how much experience each individual game might give a player. In doing so, Thompson showed that most players had a primary race, meaning they would play one of the three races far more often than the other two. Thompson identified race as a prominent nuisance variable when considering the effect of experience on FAL. Although no formal analysis compared main race games to games in which players selected a non-primary race, graphical representation showed the possibility of slower median FALs for non-primary race games. This idea is in line with previous theories of shifts in performance due to changing strategy, specifically Gray and Lindstedt's (2016) claims. Each of the three races has differences in types of buildings, military units, and ultimately strategies. Within a player's main race, any performance changes due to shifts in strategy should be apparent with faster or slower FALs. With this detailed performance measure, StarCraft 2 is a strong method for testing whether strategy-based piecewise curves can account for plateaus and improper aggregation in learning curves for a complex and natural task.

1.4. Issues With Separation of Learning Curves

The complicated methods and change point analyses of the plateau literature (Gray & Lindstedt, 2016; Rahman & Gray, 2020) break down for overlapping strategies. If the player switched between strategies often, the resulting shifts in skill would not have detectable points of change. In the plateau studies using forms of change point analyses (Gray & Lindstedt, 2016; Rahman & Gray, 2020), the explained strategies were discovered sequentially by participants. In the games used, there was a score counter visible to the participants at all times. The games also did not have dynamic opponents (the opponent was a computer program with very limited behaviour), causing high consistency between matches. The participants were able to quickly see if a new strategy was more effective than the current one. Because of this, the periods of trial and error for each strategy were considerably short. Rahman and Gray's participants made quick changes to new, more optimized strategies, resulting in distinct dips and leaps in performance. With strategies available that resulted in clear higher performance (and visible metrics throughout the game for the participants to confirm), participants did not return to prior strategies.

With highly overlapping strategies, a change point analysis would not detect the same distinct dips and leaps. Therefore, the simpler methods from the piecewise curve literature (Delaney et al., 1998; Donner & Hardy, 2015) provide a better option for capturing performance changes in StarCraft 2 strategies. Modeling separate curves for each strategy will capture the differences in optimization over time.

Due to the indirect feedback structure and high variability of opponent strategy and skill in StarCraft 2 games, players switch back and forth between strategies. StarCraft 2 does not offer an overall score or any general metrics in game for players to see how well a strategy is working. The player must rely on scouting their opponent's actions to determine their own strategy's viability for that game. However, scouting is a heuristic method that relies on the player's knowledge of the race they are playing, their opponent's race, and the game in general. This method is still prone to error as a player might think they are producing faster than their opponent and misjudge the outcome of a battle. Players can only know for certain if their strategy worked once they have won or lost. Yet, due to the matchmaking system, an opponent could have been stronger or weaker than the player which can sometimes lead optimal strategies to result in a loss and suboptimal strategies to result in a win. After games, there are some metrics such as the actions per minute or the number of workers trained which allow for players to more accurately compare between their own games. Even with these metrics, players need to use a strategy for a considerable number of games before they can decide if a strategy is viable in certain situations.

1.5. Importance of Strategy in StarCraft 2

Strategies and build orders in StarCraft 2 can influence the style and the pace of the game. For instance, rush strategies require a quick devotion of resources to produce a few military units and overwhelm the opponent before they can match military units. Rush strategies require a different allocation of resources, different build order, and different unit priority and timing than other, longer-term strategies. Comparatively, a Terran marine ball requires many barracks to create a large group of marines that can storm the opponent base and overpower them.

A strong strategy balances resources with quick early developments to 'out-build' the opponent. Each player starts with a limited amount of resources, but more of these resources can be gathered by producing more workers. As more time passes throughout the game, players produce enough workers for the resources to become abundant. Because of this increase in availability of resources throughout the game, the early buildings have a relatively strong significance to the strategy. Proper timing and decision making in the beginning of the game, when resources are scarce, can give a player an advantage for early battles. Therefore, it is beneficial for players to have the first few buildings of their strategy memorized so that they can allocate more attention to worker management, scouting, and monitoring resources to produce the buildings as fast as possible.

In order for a player to master these strategies, they must understand the relative strengths and weaknesses of the buildings and units they are producing at various points throughout the game. This knowledge allows players to capitalize on early, middle, or late game advantages that their strategy might influence. For novice players, optimal performance would require strategy-specific attention as the player learns the strengths and weaknesses of each strategy. While a player is learning a strategy, any attention and time spent on deciding which buildings to create takes away from other aspects of resource management. Therefore, the importance of build order preparation and knowledge is apparent early in the game and can give one player a possible advantage over another.

Chapter 2.

Methods

2.1. Defining Strategy

Due to the complex nature of StarCraft 2 build orders, strategy can be difficult to define in a research context. Ballinger et al. (2013), used boosted decision trees to predict player IDs from build order. Their models used each player's units trained, buildings created, and the timing of each action to determine player strategy. Then through actions per minute and strategy over time, they were able to correctly identify the player 75% of the time. These models also offered high accuracy in predicting the player's next action at 50, 150, and 250 actions into the game. The models identified influential points in classifying player strategy dependent on the units or buildings produced by certain time points. However, the researchers only used professional StarCraft 2 players in their dataset. Because of vast differences in player speed across skill levels, these models of strategy would likely not generalize to novice players. Using time dependent methods for identifying strategy might increase accuracy for expert players which show high speed and consistency, yet will likely have too much noise and variance among all skill levels to determine strategy with the same degree of accuracy.

In an attempt to simplify the methods of Ballinger et al. (2013), hierarchical clustering was considered to group games by similarity in build orders and unit production. Multiple types of list difference calculations were used to compare the build order and all units produced in each game against every other game played by a participant, leading to a different distance matrix for each player. A separate hierarchical clustering analysis was then used for each player to determine the best groupings for all of that player's games. This method of grouping games did not yield meaningful splits in the data. For instance, games in which the player clearly attempted a Terran marine ball were spread out across various different groups and mixed with other clear strategies. There was too much variability in the units produced for any of the list distance methods to consistently identify relevant strategies. The lists were then limited to just the buildings produced. These lists, although better, still did not yield meaningful splits in the clustering analysis. Even though the earlier buildings have more significance to the strategy, all buildings were considered with the same weight under the list distance calculations. Since players adapt to their opponents to a certain degree throughout the game, similar starting builds were placed in different groups due to similarities later in the game. Because the distances between every game were considerably large, it was determined that there was still too much variability in the full build orders for meaningful groupings to be made.

StarCraft 2 coach and instructor Cody “Osiris” Collins (2013) suggests that in scouting, it is most important to watch the opponent’s second unit producing building. This allows for the player to have a decent idea of the intended military composition in which the opponent is setting up for. However, strictly unit producing buildings leave out technology upgrades. For some races, technology is more important than others. Zerg, for instance, uses their main base for the majority of unit production. So, most early buildings from Zerg players are to create technology upgrades which are vital for various rush strategies. Because of the importance of technology in determining certain types of player strategies (particularly rushes), both unit producing buildings and technology buildings were included. Despite this difference in technology importance across races, there is little difference in the time required to complete various buildings for each of the three races. The main base needs 100 seconds to finish, and the early buildings are all similar for all three races (ranging from 45 to 65 seconds). There was also no evidence for race effects in the time it took for players to complete the first three buildings (see Appendix A).

Determining the number of buildings to use for strategies presented some difficulty. There was a tradeoff of information and variability depending on the number of buildings selected for each game. Osiris suggested the first two buildings were the most useful in scouting yet including three or four buildings in the analysis would allow for even closer discrimination between various army compositions and means of attack, allowing a more descriptive definition of strategy. However, the tradeoff for using three or four buildings is the increased variability of player strategies. The number of possible builds exponentially increases in a tree-like fashion for each extra building included. Because of this, 3 buildings seemed to yield the most information while still keeping a reasonable average number of games per strategy. With the reduction in variability between games by limiting the build order to the first 3 buildings, no clustering analyses were needed. Instead, games were simply grouped categorically by the first 3 buildings.

2.2. Data

The data used for the present study were originally collected and used in the Thompson (2018) dissertation. The study was advertised on various online forums which were popular for StarCraft 2. Players filled out an online survey and uploaded all the game replay files on their account. Data were collected in 2014 and included games ranging back to the release of StarCraft 2 in 2010. Participants were advised to submit all the game files played from their account. There were 119,334 total games collected after basic validity checks. Replay files were parsed using SC2Gears (Belicza, 2014), then organized and stored in a MySQL database. There were two participants that reported sharing their account, but they stated that this was a very rare occurrence. Of the hundreds of games from these players, there was no discernable pattern to tell which games were played by someone else. All their games were included. There was one major update to the game (Heart of the Swarm) in 2013. The dataset includes games both before and after this update. Several of the game military units were changed, but an analysis from a different study on this same dataset (yet to be published) has suggested there is no sufficient change in FAL before or after the update, so no further influences of game updates were considered.

In a typical learning study, games would be ranked by time and the rank numbers would be used in the model. However, there are gaps in play for most of the players. Many players have groups of games played back to back and gaps of various amounts of time with no games. These back to back games and gaps could explain dips and leaps in performance, which would not be factored out with the game rank numbers. In order to account for the time between games, the number of days since each player's first game was used as a predictor in all models tested. This variable not only accounts for clusters of games played back to back, but also takes into account any breaks or gaps in time the player might have taken in between games. However, there are two types of gaps that might occur in the data: true gaps (where a player completely takes a break from the game) and gaps in competitive play (where a player might have some games in alternative game modes). The gaps in competitive play led to a complicated relationship with experience (Thompson, 2018). In the interest of keeping as simple of a time variable as possible, non-competitive experience was not included. Therefore, each player's first game after all exclusions was used as day one.

The existence of failed buildings within the tables caused some issues with the strategy variable. If the production of a building was interrupted at any point, the build command was coded as a fail. This includes the worker temporarily being directed away for other actions, the worker getting attacked, or the partially completed building getting destroyed by the enemy. If the worker was moved away or attacked, the dataset did not include a reliable way to determine if a worker was moved back to finish producing the building. If all failed buildings were excluded, this would throw out a large number of legitimate buildings. If all failed buildings were included, some buildings that were never completed might be included in the build order. Yet even if the building fails, the resources still get spent. Because of the importance of resource management in StarCraft 2, it is very unlikely that players would leave an unfinished building in their base. Using the resources for a building early in the game still shows the intent and commitment of a specific strategy, whether or not that building ends up fully completed. For the purpose of more accurately capturing the players' intent in the strategy variable, all failed buildings were included.

2.3. Analyses

Typically, researchers use mixed effects models to fit learning curves. However, in the current dataset there are many possible build orders for each of the three races. Because of the variety of strategies, a single model fitting everyone would need a new component for each strategy that was used. This would result in a model with 28 strategy parameters (56 when including the interaction effects). The following likelihood ratio tests would have an unnecessarily inflated effect size in accounting for all the possible strategies together. Weighted cross validation and alternatives to the likelihood ratio test were considered to help in estimating a more accurate test statistic and controlling for differences in number of games between participants, yet a preliminary power analysis conducted with the simr package (Green & MacLeod, 2016) yielded 100% power in all categories. This analysis included a condition with no effect for any strategy and similar variance in the simulated data as observed in the present dataset. Considering there was still essentially 100% power even with a null effect as long as there was random noise in the data (showing highly inflated type 1 error rates), other options for modeling were considered.

Instead of using a single model to fit all participants, a more exploratory method using separate models for each participant and meta analytic methods to summarize effects allow for a less biased estimation of the general effect of strategy on learning curves. Each individual model would only have to fit parameters for the strategies that the player used. In order to further limit model complexity, only the three most common strategies for each player were used. This reduction in dimensions for each model allowed more accuracy in the model comparison tests, which were then summarized using meta analytic tests for significance in average effect sizes.

Initially quadratic curves were considered for fitting these individual models. This method would have allowed a conceptually and graphically easier representation of the learning curves than the log-log models typical in power law research. However, fitting a quadratic model requires an additional component for the squared effect on time as well as an extra component for each strategy in the interaction effects. The increased complexity of these models did not provide a visually better fit compared to the log-log plots. Therefore, the log-log method was chosen for its relative simplicity.

For each participant, three nested linear models following the log-log-linear method (Newell & Rosenbloom, 1981) to predict the log values of FAL were used. The simplest model only contained a log transformed predictor of time (days played, starting from 1). The next model had the first time component and an effect for categorical strategy which contained two or three levels depending on the player. The final and most complex model added an interaction effect between strategy and days played. Models were compared by the change in R^2. Total effect size for the strategy main and interaction effects were determined with Z tests on the change in R^2 between the models weighted by inverse squared standard error. These methods are consistent with recommended meta analysis techniques (Card & Little, 2012). Standard errors for the R^2 values were computed using a formula from a multiple regression textbook (Cohen, Cohen, & West, 2014).

2.4. Exclusions

StarCraft 2 has multiple game modes that a player can select. There are 1 versus 1 competitive games as well as team games with more than one player on each side. In competitive games, wins and losses determine rank. Players generally consider competitive to be the primary game mode. For the purpose of this study, only interested in 1 versus 1 competitive games are of interest. This exclusion criteria results in a higher likelihood that each game is taken seriously by the participants, while also dropping any different strategies and playstyles used in the team games. 35,923 games were excluded for not being competitive 1v1 matches. 766 more games were excluded for having total lengths of under 20 reaction times in total. An additional 5,576 games were dropped because the player or their opponent did not create at least 3 unit producing buildings.

In-game race is a significant nuisance variable (Thompson et al., 2019 significance test for race). Therefore, only players' main race (defined as the most common race they play) games were used in order to avoid unnecessary within-player variance. 4,309 non main race games were dropped.

By including failed buildings, there were some issues in the data with impossible building locations. If an enemy put a unit in the space where the player was trying to put down a building, the build action was canceled, and this was coded as a failed building. Because of this, 2 or more build actions (failed or successful) could occur on the same location. In games where this occurred, it was difficult to tell which of the build commands, if any, actually resulted in the building finishing production. Therefore, any games which had two or more of the first three buildings on overlapping coordinates were excluded, resulting in 3,145 games getting dropped. However, it is possible for players using the Terran race to lift and move their buildings once they are completed. So for Terran players, overlapping coordinates are possible as long as there is adequate time for the building to be completed and moved. This time was chosen as 50 seconds since this is the lowest completion time for any of the Terran buildings included. To account for the possibility that Terran players moved buildings, 595 games that included Terran buildings with overlapping coordinates, but were at least 50 seconds apart were re-included. In summary, a total of 2,550 games were dropped for containing impossible building locations.

Additional games were dropped using the parameters described in the power description. Limiting the players to three strategies and only including strategies with 30 games each resulted in 28,622 games being dropped.

There were large gaps in the resulting timelines for some players. Multiple players only had a few games before taking months off, or would take several months break with only a few games at the end of their dataset. In order to prevent these few games from having disproportionate leverage on the models, games were only included if they were within the first and last points where participants played at least 10 games in 30 days. 172 games were excluded as outliers in days played.

There was only one game which was a notable outlier in FAL. This game was excluded as well because the average FAL was over one minute, suggesting the player was not paying attention to their computer for at least part of the game. The resulting dataset had 41,415 games across 76 players. There were 66 players with three strategies included and 10 players with two. There were 23 Protoss, 20 Terran, and 33 Zerg players remaining. There were 9 overall strategies for Protoss, 7 for Terran, and 13 for Zerg.

2.4.1. Power

Harrell (2015) suggests $p < m/15$ where p is the number of parameters and m is the "limiting sample size" for small R^2. Each of the larger models will have 5 parameters for players with 3 strategies (log(days), strat2, strat3, log(days):strat2, log(days):strat3) or 3 parameters for players with only 2 strategies. For players with 3 strategies, over 75 should be used ($5 < m/15$; $m > 75$). For players with 2 strategies, over 45 games should be used ($3<m/15$; $m > 45$). Within each categorical strategy, there would be 2 variables fit (main effect and interaction), which suggests over 30 games per strategy should be used ($2 < m/15$; $m > 30$). With the above conditions as a ballpark estimate for the number of games per strategy to include, over 30 games per strategy would meet all 3 of the above recommendations as well as the recommendation for at least 60 degrees of freedom for the standard error of R^2 calculation (Cohen, Cohen, & West, 2014). Therefore, the top 3 strategies per player were selected. Only strategies with over 30 games and players with more than one strategy were included.

2.5. Demographics

Of the 76 players included in the analysis, 2 were missing survey data, and 1 did not answer any of the questions. There were 70 participants that identified as male, 2 as female, and 1 as other. The most common country inhabited was the United States (21), with Canada (7) as the second most common. All other country information is in Table B.1. The mean age of participants at the time of data collection was 26, with a minimum of 18 and maximum of 40. The players mostly had some sort of higher education: 15 reported highest as middle or high school, 19 reported an associate degree or college diploma, 27 reported university bachelor's degrees, and 12 reported graduate school. A majority of the participants stated that they watched StarCraft 2 games online (69), 3 players stated that they did not, and 1 did not answer. Most of the players had played StarCraft 1 (48; 24 never played StarCraft 1 and 1 did not answer). Because of the lack of non-male participants, no generalizations will be made for other genders.

Chapter 3.

Results

Three models were fit for each participant using R (R Core Team, 2012), then the weighted averages were taken for the change in R^2 between each model. The base model regressed log (base 10) transformed Days Played onto log (base 10) transformed FAL. Through taking the weighted average of the parameters across participants, this model yielded an intercept of 2.874 ± 0.001 and a slope of -0.057 ± 0.0006. The base model was a significant fit ($Z=16.187$, $p<0.001$) with a weighted mean R^2 of 0.011 (95% CI [0.009, 0.012]). The individual effect sizes for each participant are listed in Figure 3.1 with a line at the mean R^2 value. Player 59 held a majority of the weight (over 67%) for the mean R^2 of this model (see Figure 3.1), despite only having 344 games after exclusions. This is due to the low R^2 and comparatively miniscule standard error for this player. This player was treated as an outlier and all analyses were run with and without including player 59. Although the R^2 of the base model increases to 0.033, there were no other changes in any of the coefficients, changes in R^2, or significance tests for any of the models. It was therefore deemed that this player did not have a sufficient influence on the analyses and was included for the following results.

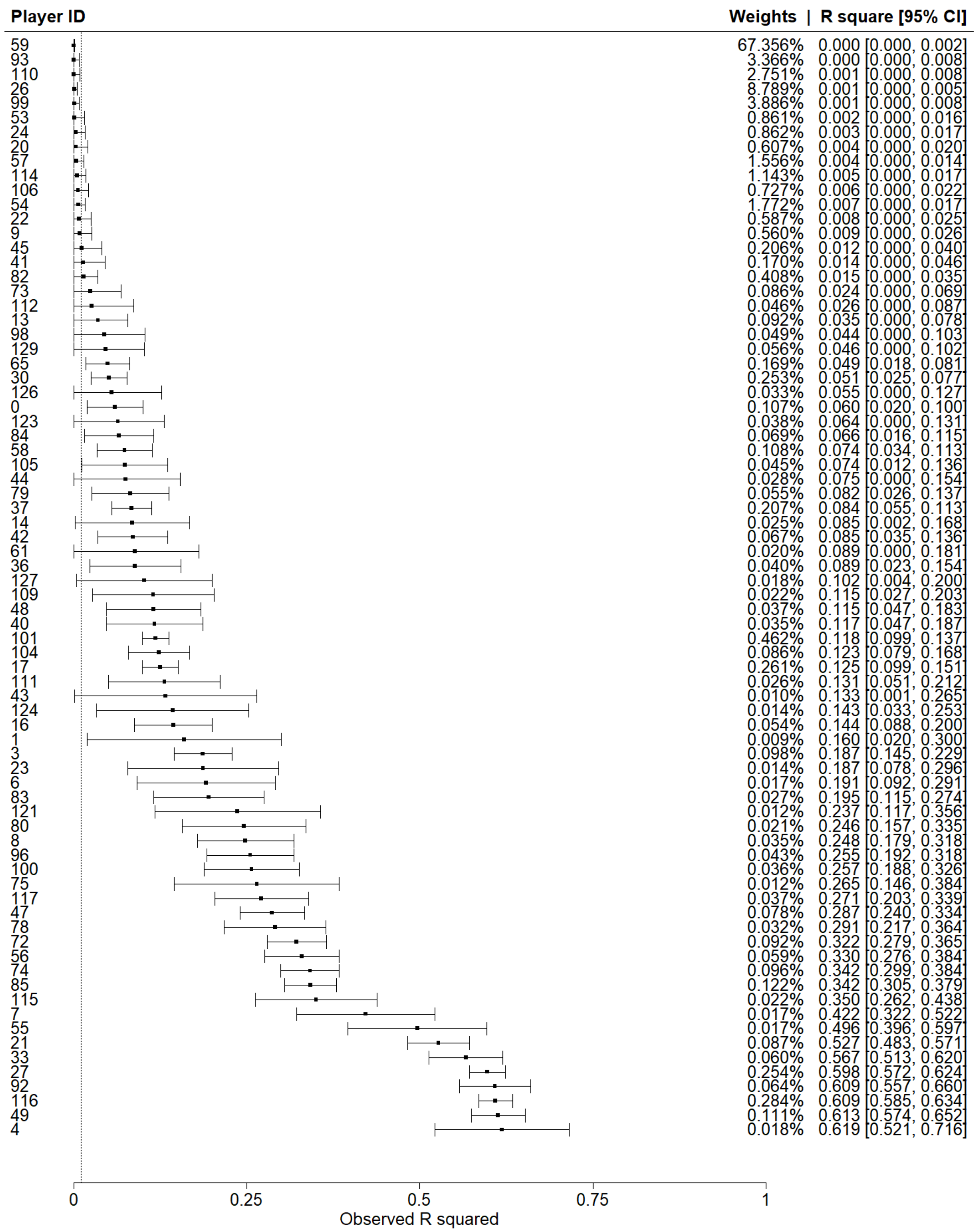

Figure 3.1. Forest plot of base model effect sizes (R^2)

The next model added the top three strategies to the model for each player. All the weighted mean beta values for each parameter are listed in Table C.1. Adding strategy resulted in a mean increase in R^2 of 0.031 (95% CI [-0.015, 0.078]) which was not a significant change (Z=1.320, p=0.093). The individual effect sizes for each participant in this model are listed in Figure 3.2 with a line at the mean change in R^2.

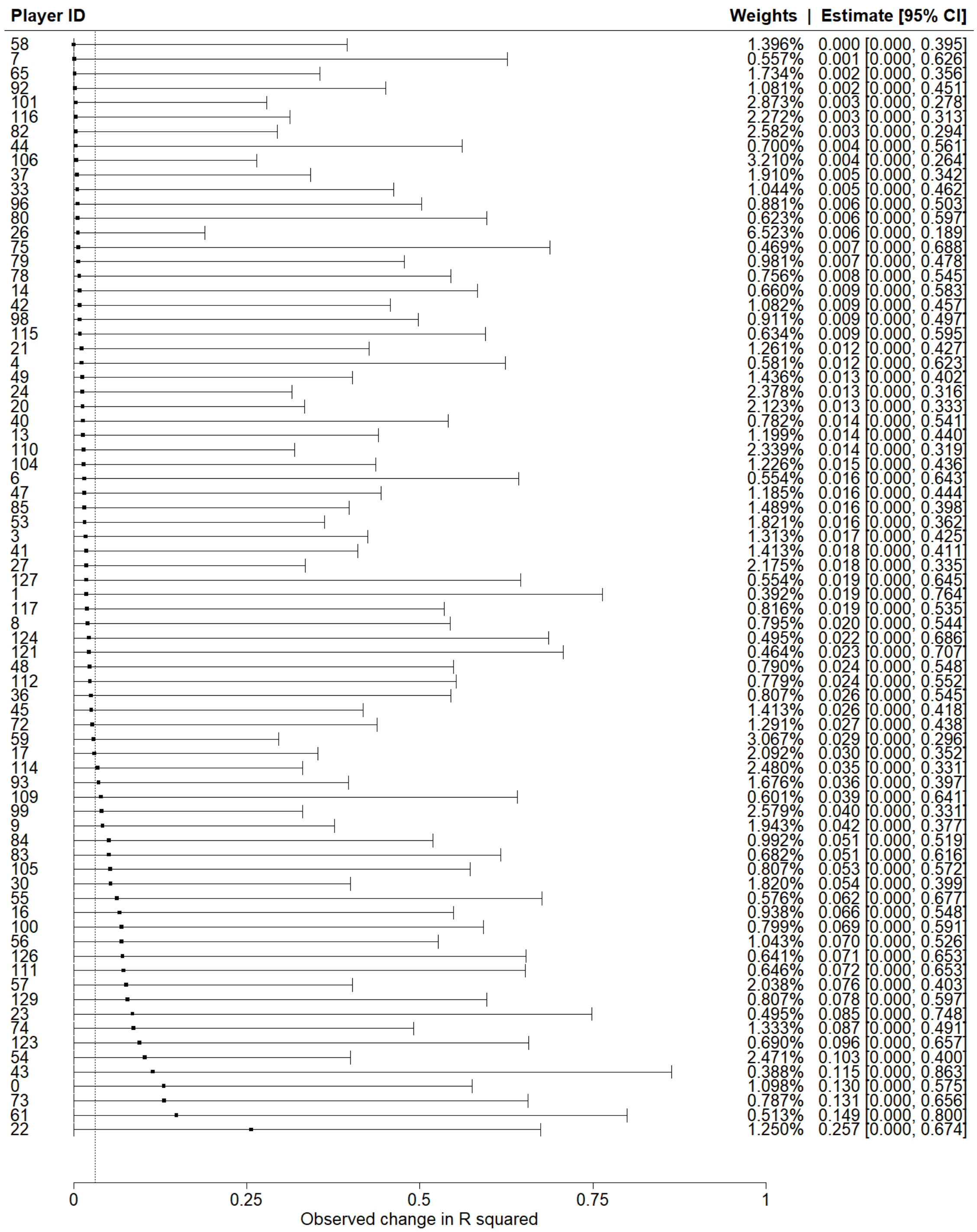

Figure 3.2. Forest Plot of effect sizes when adding strategy (ΔR^2)

Finally, the third model added an interaction term between strategy and days played. The weighted mean beta values for each of these parameters are listed in Table C.2. The addition of an interaction term did not significantly increase the mean R^2 ($Z=0.479$, $p=0.316$; $\Delta R^2=0.013$, 95% CI [-0.039, 0.064]) compared to the second model. The individual effect sizes for this model are listed in Figure 3.3 with a line at the mean change in R^2.

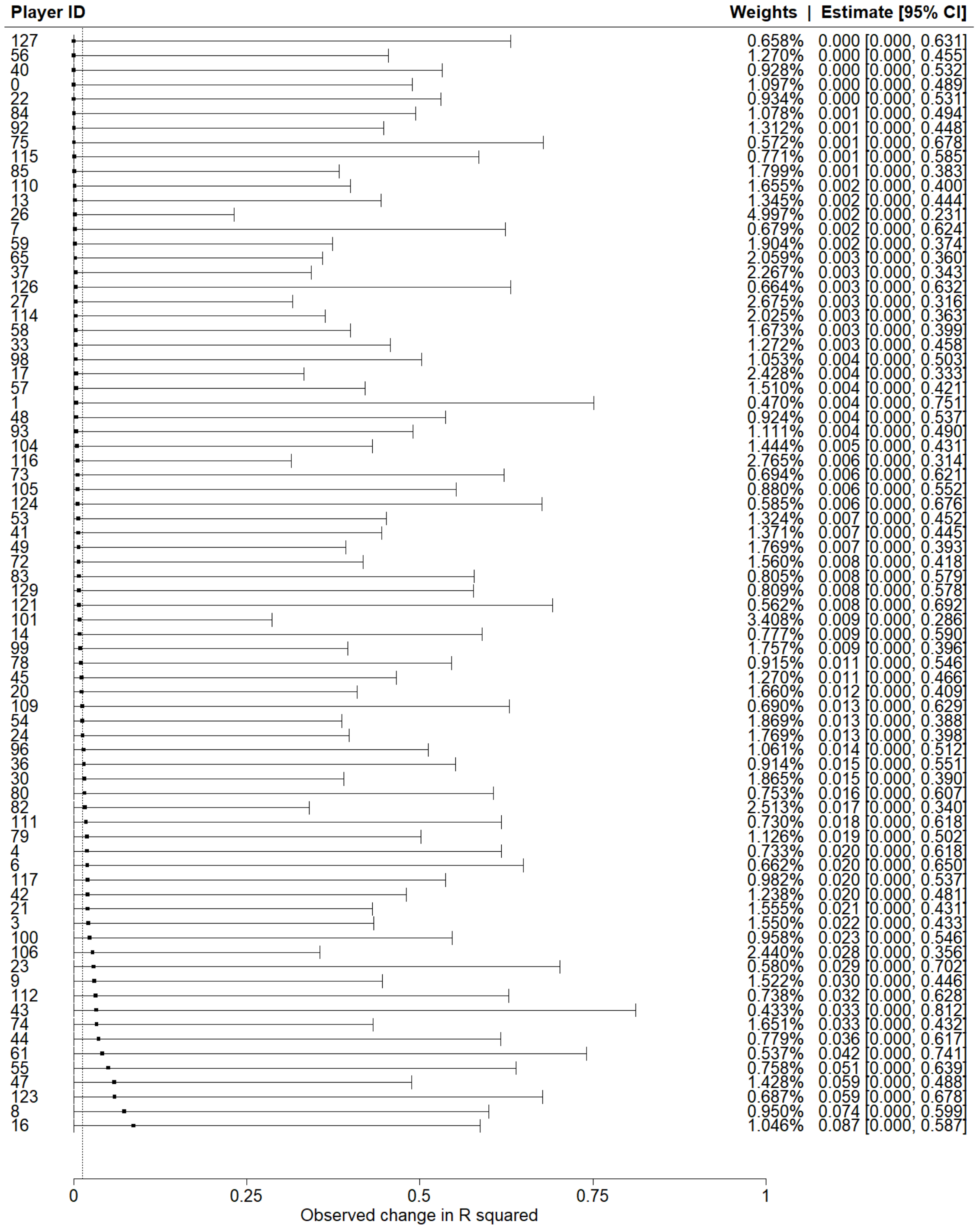

Figure 3.3. Forest Plot of effect sizes when adding an interaction (ΔR^2)

The small insignificant changes in model fit by including strategy and the interaction between strategy and days played can be seen in Figures 3.2 and 3.3 where many players had little to no changes between the models. Figure 3.4 shows this pattern more clearly by plotting the distribution densities for each of the three model R^2 values.

Although there are slight differences between the models, the distributions still have a high degree of overlap.

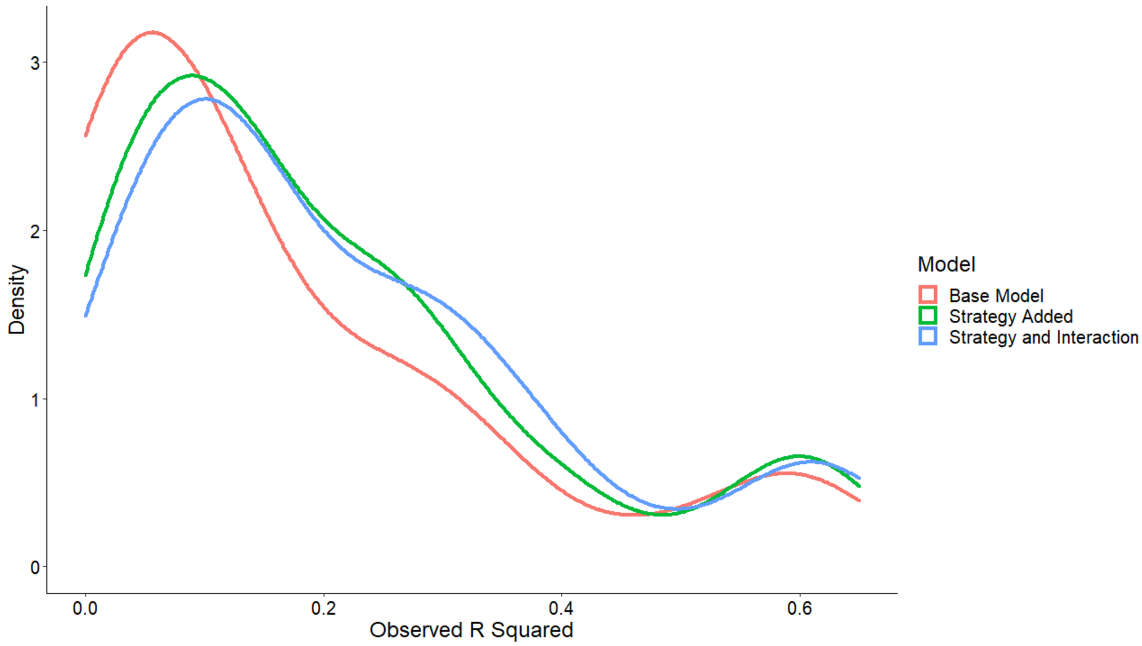

Figure 3.4. Density plot of R^2 values

An example of a typical participant can be seen in Figure 3.5: a scatterplot of the data from player 9, who had R^2 similar to the mean values for each model. The black line in Figure 3.5 is the best linear fit for the base model, while the coloured lines are the best fit for each strategy with standard error shaded regions. There is a high degree of overlap between all the error regions and the strategies seem to only be slightly different from one another. Player 9 also had a large gap in which they did not use the "Barracks, Command Center, Factory" strategy which could result in high leverage for the first game with that strategy. This pattern of gaps within strategies was relatively common among the players used in the present study. It is also shown in Figure 3.5 that player 9 did not actually get faster at the game, despite playing 438 games across 49 days. Many of the players in this dataset did not improve or only did so marginally. In fact, only 59% of the players (45 of 76) made visible improvements in FAL. 21 of those players (28% of total) appeared to improve along any kind of a curve.

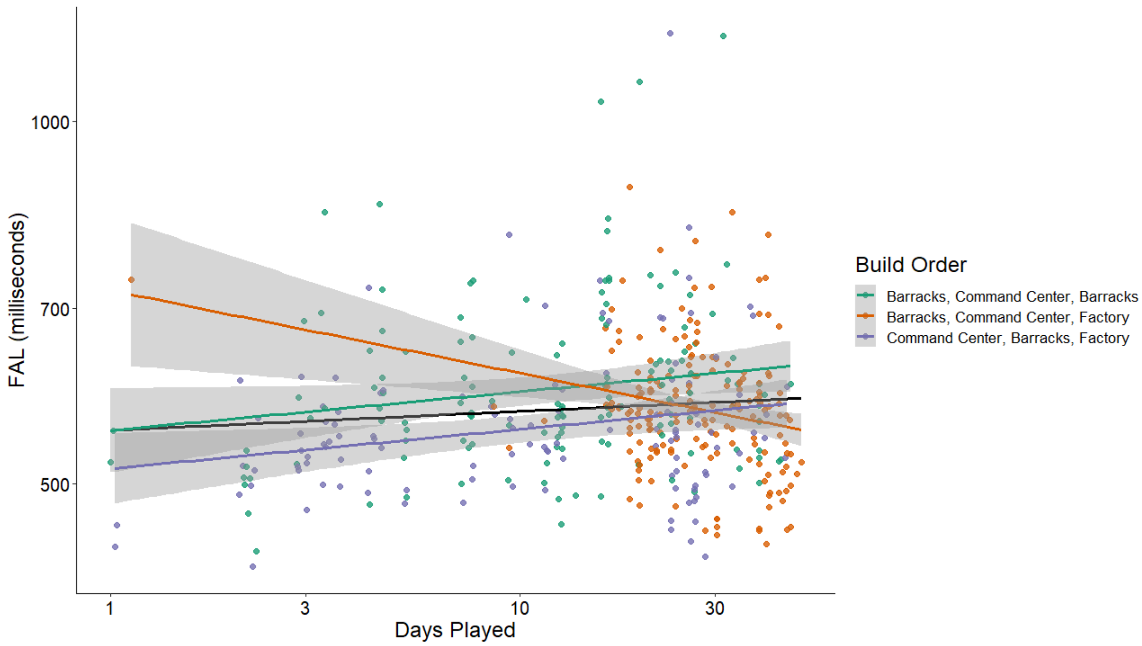

Figure 3.5. Log-log scatterplot for Player 9

Age was considered as a possible confounding variable for the lack of smooth power law learning curves. Prior research has shown aging effects on StarCraft 2 FAL starting at age 24 (Thompson, Blair, & Henrey, 2014). Considering 41 of the players in the present dataset were 24 or older at the time of submitting their game files, this could offer a reasonable explanation. However, the data only range from 2010 to 2014, and the longest that any player included in the present study played was 2.8 years. Considering the effect size listed in the supplementary information of Thompson, Blair, and Henrey's paper was a reduction in log FAL of 0.008 per year of age after 24, this should not explain the lack of learning in the data. A theoretical maximum aging effect of 0.024 (0.008 across 3 years) is not a large effect in comparison to the average rate of learning found in the base model (β= -0.057 ± 0.0006). Therefore, it is not likely that aging would sufficiently explain the large number of non-learners or the lack of power law curves in general. Nonetheless, a weighted regression was performed for each model as a double check predicting R^2 by player age. Aging did not result in any significant impacts on R^2 for any of the models.

Chapter 4.

Discussion

Skill over time is thought to follow an approximate power law curve (Newell & Rosenbloom, 1981). Three areas where the power law curve breaks down were identified: there are strategy-based inconsistencies in learning curves that can be attributed to piecewise curves (Delaney et al., 1998; Donner & Hardy, 2015; Rickard, 1997), improper aggregation (Gallistel, Fairhurst, & Balsam, 2004; Haider & Frensch, 2002; Myung, Kim, & Pitt, 2000), and plateaus (Bryan & Harter, 1897; Bryan & Harter, 1899; Gray & Lindstedt, 2016; Rahman & Gray, 2020). The shifts away from typical learning curves had been shown in simpler lab studies (Delaney et al., 1998; Donner & Hardy, 2015; Haider & Frensch, 2002), but were also observed in more complex tasks (Bryan & Harter, 1897; Bryan & Harter, 1899; Gray & Lindstedt, 2016; Rahman & Gray, 2020). The present study worked to explain inconsistencies in the power law of practice by attempting to generalize strategy-based shifts in skill to a highly complex task. However, an important underlying assumption of the strategy-dependent learning curves is that participants learn the task and follow a general curve. This was not always the case for StarCraft 2 players.

4.1. The Standard Learning Curve

Complex tasks yield complex data. The initial models following the standard log-log-linear method (Newell & Rosenbloom, 1981) showed a poor fit to the present StarCraft 2 data. The expectation of prior learning research is that practice leads to improvement. Every article on learning cited in the present study has an increase in skill over time. The only extended period of non-learning was observed by Bryan and Harter (1897; 1899). They found a large plateau in telegraph operators. Further plateau research associated strategy with periods of non-learning (Gray & Lindstedt, 2016; Rahman & Gray, 2020). Yet even in the plateau literature, participants made overall improvements over time. There were no participants listed that completely failed to get better at the task.

The present dataset did not match the learning curves predicted by the power law of practice. Although the power law theory described by Newell and Rosenbloom (1981) has been widely accepted in learning research (Anderson, 1982; Ericsson, Krampe, & Tesch-Römer, 1993; Logan, 1988), the StarCraft 2 data presented in this study does not match the standard log-log-linear method to model learning curves. 45 players (59% of the total) improved throughout their submitted games, while only 21 of those players (28% of the total) showed any resemblance to a curved learning pattern. This study is unable to support the initial claims towards the ubiquity of the power law of practice. Much like the data from Bryan and Harter's research (1897; 1899), the power law does not always apply to large scale observational data.

4.2. The Effect of Strategy

Bryan and Harter suggest shifts in participants' understanding of the task might account for the observed plateau. This idea that shifts in understanding could influence learning curves led to the investigation of strategy. Shifts in strategy accounted for plateaus within participants' learning curves (Gray & Lindstedt, 2016; Rahman & Gray, 2020). Therefore, providing there are still similar shifts in understanding of the task, strategy should still have an observable effect for players that did not improve over time. A model that includes strategy should still help to explain the poor fit of the general learning curve, despite the novelty of the present dataset. Yet the added specificity of strategy did not help to improve models. Many players failed to improve after hundreds or even thousands of games, but also failed to improve across the same strategy: playing the same way and performing similar actions across hundreds of games.

Despite clear strategy-dependent learning curves in the simple arithmetic and logic tasks of the piecewise curve literature (Delaney et al., 1998; Donner & Hardy, 2015; Rickard, 1997), the present study has failed to generalize this trend to StarCraft 2. It is possible that the broader definition of strategy used in the present study does not accurately capture the changes in understanding of the game experienced by the players, like the strategy described by Delaney et al. (1998). Another explanation would be the strategy-dependent curves do not generalize to tasks as complex as StarCraft 2. Delaney et al. (1998) and Donner and Hardy (2015) used very simple learning tasks and found clear increases in R^2 for their piecewise curve models.

The plateau literature however, expanded the definition of strategy and increased task complexity, while still finding similar leaps followed by plateaus with each strategy (Gray & Lindstedt, 2016; Rahman & Gray, 2020). The strategy-dependent plateau trends have also failed to generalize to StarCraft 2. Plateaus from different strategies would have been detected in the test for the main effect of strategy. Although strategy-dependent plateaus were not detected in the present study, none of the change point analyses or more complex statistical methods developed by Rahman and Gray (2020) were used for the present data. Therefore, the results of this paper do not refute, nor confirm the general existence of plateaus in observational learning data (Bryan & Harter, 1897; Bryan & Harter, 1899; Gray & Lindstedt, 2016; Keller, 1958). Yet, the methodological developments from this paper offer a feasible starting point for other studies to test the generalization of strategy-based curves in other complex tasks.

There may be other variables to control when considering observational learning data for highly complex tasks, especially when the data include novices or potentially casual players. The resulting low R^2 from the base log-log-linear set of models along with several participants whose FALs did not improve over time (see Figure 5) suggests that there may be some unmeasured variables which may have influenced the player's learning abilities. Mere experience does not always guarantee that a player will improve at the game. There are other factors that could influence the efficiency of practice: effort, duration of practice sessions, and motivation (Ericsson, Krampe, & Tesch-Römer, 1993); the consistency of practice sessions (Huang et al., 2017); or sleep (Born, Rasch, & Gais, 2006). Ericsson, Krampe, and Tesch-Römer outline that mere practice is not enough to improve at a task. The practice must be deliberate for someone to get better. They describe the effect of effort in practice: an explicit goal to improve, the active seeking of more optimal methods, and the motivation to improve performance. It is possible that the act of switching strategies is not enough on its own to improve performance, but only when players make a meaningful and intentional switch in strategies in an effort to improve a certain aspect of their skill.

4.3. The Hobbyist

If deliberate practice and the effort that people are willing to give are the main areas that separate amateurs from experts, then there should be a group within most observational studies of all the people who do not care if they improve. This group can be deemed the hobbyist population: those who are content with non-expert performance. The hobbyist group is what sets apart observational studies of complex tasks from simple tasks or retrospective analyses of professionals. Study two of Ericsson, Krampe, and Tesch-Römer (1993) considers the practice times of amateur pianists. They suggest that the overall accumulated practice time for amateurs is far less than that of experts, with deliberate practice being the main difference between the two groups. Ericsson, Krampe, and Tesch-Römer offer one of few studies to consider the hobbyist group.

In the present study non-learners were not excluded, leaving all possible hobbyists included in the analyses. If it is assumed that all players who did not improve were just hobbyists, then there should be common themes between the intent of the non-learners. Participants in the present study had the option to leave a comment with their data submission, in which some players listed general comments that may have impacted their performance. Of the players that did not improve (and chose to leave comments about their play; n=8), there were identifiable general themes: low motivation (one player), lack of time (due to school, work, or interest in other games; two players), or lack of domain specific knowledge (they did not know what parts of their performance to focus on in order to improve; one player). These three themes might be useful in identifying hobbyists in other research, yet there were some inconsistencies in responses.

There are still potential hobbyists that improve over time. Two players listed low time but still managed to improve at the game. This may speak to the effort of those players' practice. If the player only has time for a few games a week, they might be putting in extra effort to make sure they improve. This potentially rules out the low time category as a consistent identifier of hobbyists.

Alternatively, there were some players that listed signs of effort in their practice but did not improve. Three of the non-learners listed explicit practice routines with time delegated for introspection, study of build orders and timings, and goals for improvement. One of the non-learners was even a professional strategist for a prominent StarCraft 2 team. These comments suggest that the hobbyist population likely does not fully overlap with the non-learners. Yet all of these effortful non-learners (including the team strategist) had high change in R^2 when including strategy or the interaction in the models. The higher effort in practice in these players might lead to separations in strategy as these players are more likely to have meaningful changes in strategy. Hobbyists might make up a larger proportion of non-learners; and paired with strategy, it is possible that this could explain the lack of smooth power law learning curves in observational data in complex tasks.

The existence of hobbyists bridges the gap between lab research with simple learning tasks and complex everyday learning. For instance, not everyone is an expert at all tasks in a daily routine. Someone could make coffee every morning and still not reach the level of a skilled barista. The intent and goals of the task matters to the learning outcomes. This hypothetical person could be content without expertly made coffee every morning, so putting in the extra time and effort to learn proper coffee making is not worth it for them. Similarly, someone could use StarCraft 2 as a way to relax after school or work. If the goal of the game is to relax or have fun, the performance outcomes measured in learning research may not always capture that.

4.4. Limitations

The current project was primarily limited in demographics and qualitative design. The demographic information previously mentioned is likely representative of the population of StarCraft 2 players at the time of data collection. However, the lack of non-male participants limits any sex or gender generalizations. Also, the presence of higher education and the need for a computer with stable internet to play the game suggests there are possible limits to the socioeconomic generalizations that can be made. This sample applies primarily to the average male video game-playing adult. Further limitations were met with the qualitative statements made when describing hobbyists. Because the only open-ended question on the survey was an optional section for comments, a few relevant replies were mentioned. Yet there was little more information that could be used from this section of the survey. Future research should consider qualitative and mixed methods research to further address some of the themes addressed in participant comments.

4.5. Directions for Future Research

Because the standard linear mixed effects models would not have produced accurate results, exploratory methods are required to research StarCraft 2 strategies.

This shift away from popular models is a sufficient setback for the practicality of generalizing strategy-dependent learning curves to other complex tasks with many possible strategies. The modeling methods developed in this paper offer a possible methodology for future studies that is simple compared to the methods of Rahman and Gray (2020). The use of individual models aggregated through meta analysis techniques provides an alternative to standard mixed effects models that can better handle high dimension strategy variables, while still preserving the ability to detect differences in strategy when participants frequently switch back and forth between their most popular strategies.

Yet with the different modeling methods presented in this study, there are still some considerations to be made in future research. There were three prominent categories of setbacks in this study that should be addressed in future research: the definition of strategy, large gaps in days played, and the lack of power law learning curves and asymptotic performance.

As discussed in the Methods section, strategy was particularly difficult to define for StarCraft 2. The use of a more detailed method capturing building timings and army composition could more thoroughly capture the intent of the player. However, any increase to the number of levels in the strategy variable leads to the need for more games per strategy. In this study a balance was found by using the simplest build order categorization that still had relevance to the game as dictated by a prominent member in the game's community. Future research should consider a task which has a clear separation of strategies that can be validated through community members. A different complex task such as chess could provide a strong definition for strategy. Chess openings, which are well rehearsed among high level players, have well documented and community-backed groupings based on the first few moves of the game. Chess is therefore an option for a different complex game with a well-defined strategy variable.

Gaps in days played created inconsistencies in the data which were difficult to account for. Large gaps can lead to forgetting, counteracting the overall learning effect. It is possible that players could forget general game knowledge and become slower after a gap in play, or they might forget strategy specific knowledge and become slower after gaps within strategy. Future research should explore the differences in forgetting between general skill and strategy specific skills. Researchers could also consider the motivational relationships for switching strategies. Player motivation and self-perception of improvement could account for some of the gaps within strategy, especially in games like StarCraft 2 with indirect feedback and high variability of opponent skill. Yet, until the reasoning behind gaps in days played is more researched and developed, future researchers might consider excluding any participants with large gaps in time (both overall and within strategy).

Much like the issues pointed out in the Keller (1958) study in the plateau section, the present analyses included many participants that did not improve along a curve and never reached asymptotic performance. All participants that did not show clear learning curves were still included in the interest of retaining as many participants as possible and reflecting practical data collection practice as it is typically difficult to collect high amounts of long-term longitudinal data. This study has shown that strategy dependent curves are not detectable under these conditions. However, future researchers should consider larger datasets so they can further restrict to participants that improve along a curve and reach asymptotic performance. Chess or other video games provide a strong option for future research because of the abundance of online datasets that include information on all games played for a large number of users. By analyzing these large public datasets, researchers could retain ample players despite any restrictive exclusions.

Although chess has no direct FAL equivalent, the Elo rating system used to rank players could be a useful alternative. General score measures that capture player skill were used by Rahman and Gray (2020) in the game Space Fortress. These scores still offered learning curves with the plateaus discussed in the introduction. Chess has a well documented international rating system that will score players based on how well they perform and the ratings of their opponents. This Elo rating has been shown to follow general learning curves over time (with some debate as to whether power law or exponential curves offer a better fit; Elo, 1978; Gaschler et al., 2014; Howard, 2014; Howard, 2018). Because the Elo ratings follow a general curve over time, these scores can be tested within participants for effects of opening strategies.

Chess would allow for consideration of all three of the above concerns, while still offering the added complexity of a human opponent and indirect feedback. Chess has relatively well-defined opening strategies that would be easier to separate than in

StarCraft 2. There are also many large datasets including games (played both online and over the board) across the lifetime of various high ranked players as well as online games from a large number of online accounts for players from a wide variety of skill levels. These large datasets would allow for further restrictions of participants to explore the nature of strategy-based learning curves under more controlled conditions. Chess also, like StarCraft 2, offers a more complex task than previous plateau research as the player must still consider and adapt to their opponent’s actions. There is already a broad range of cognitive science research focused on the strategy, planning, and pattern recognition for chess which could be considered moving forwards (Charness et al., 2001; Gobet, 2010; Saariluoma & Hohlfeld, 1994; Van Der Maas & Wagenmakers, 2005). There is even already some research on the opening move orders (Blasius & Tönjes, 2009; Cero & Falligant, 2019). Yet there is still a gap in the literature specifically analyzing the changing of strategies and how those shifts could affect performance. Prior research has focused on finding consistencies between players’ strategies. Now there is a research base to be able to determine how the players might use certain strategies to improve. For the reasons stated above, chess would be a useful task for future research.

Chapter 5.

Conclusion

The findings of this study suggest that the power law of practice may not be as universally found as once thought. Considering the poor initial fits of the log-log-linear method of analysis, it is possible that observational data of a variety of skill levels in highly complex tasks does not always present a learning curve. Strategy-dependent learning curves did not provide any substantial improvements to the power law models or help to explain the lack of traditional learning curves. The concept of hobbyists (people that do not intend to reach expertise) was introduced as a possible explanation for the lack of power law learning curves in the dataset. Further consideration of hobbyists might help to account for lacking power laws in observational data of complex tasks. There are still considerations to be made for future research before strategy is completely ruled out as an important effect. The meta analytic style methods of this paper should be further considered for testing strategies of complex tasks: offering higher accuracy than mixed effects models, which would have detected a false positive effect. The analyses used in this paper also allow for a structured inclusion of strategy and relatively simple calculations compared to the complex methods of the plateau literature. Further guidelines were specified for future research, with chess appearing to be a strong option for the next study of strategy in a complex task.

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369–406. doi:10.1037/0033-295X.89.4.369

Ballinger, C., Liu, S., & Louis, S. J. (2015, October). Identifying Players and Predicting Actions from RTS Game Replays. In *Proceedings of the 28th International Conference on Computer Applications in Industry and Engineering, San Diego, CA, USA* (pp. 12-14).

Belicza, A. (2014). StarCraft 2 Gears (14.3.3) [Computer software]. Retrieved from https://sites.google.com/site/sc2gears/

Blasius, B., & Tönjes, R. (2009). Zipf's law in the popularity distribution of chess openings. *Physical Review Letters*, *103*(21), 218701. doi:10.1103/PhysRevLett.103.218701

Born, J., Rasch, B., & Gais, S. (2006). Sleep to remember. *The Neuroscientist*, *12*(5), 410-424. doi:10.1177/1073858406292647

Bryan, W. L., & Harter, N. (1897). Studies in the physiology and psychology of the telegraphic language. *Psychological Review*, *4*(1), 27. doi:10.1037/h0073806

Bryan, W. L., & Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological review*, *6*(4), 345. doi:10.1037/h0073117

Card, & Little, T. D. (2012). *Applied meta-analysis for social science research*. Guilford Press.

Cero, I., & Falligant, J. M. (2020). Application of the generalized matching law to chess openings: A gambit analysis. *Journal of applied behavior analysis*, *53*(2), 835-845. doi:10.1002/jaba.612

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory & cognition*, *29*(8), 1146-1152. doi:10.3758/BF03206384

Cohen, P., Cohen, P., & West, S. G. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences*. Taylor & Francis Group.

Collins, C. (2013). *Scouting in Starcraft 2*. Osiris Starcraft 2 Guide. Retrieved from https://osirissc2guide.com/scouting-in-starcraft-2.html

Compton, B. J., & Logan, G. D. (1991). The transition from algorithm to retrieval in memory-based theories of automaticity. *Memory & cognition*, *19*(2), 151-158. doi:10.3758/BF03197111

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological science*, *9*(1), 1-7. doi:10.1111/1467-9280.00001

Donner, Y., & Hardy, J. L. (2015). Piecewise power laws in individual learning curves. *Psychonomic Bulletin & Review*, *22*(5), 1308-1319. doi:10.3758/s13423-015-0811-x

Elo, A. E. (1978). *The Rating of Chessplayers Past and Present*. American Chess Foundation.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. doi:10.1037/0033-295x.100.3.363

Gobet, F. (1997). A pattern-recognition theory of search in expert problem solving. *Thinking & Reasoning*, *3*(4), 291-313. doi:10.1080/135467897394301

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*(36), 13124-13131. doi:10.1073/pnas.0404965101

Gaschler, R., Progscha, J., Smallbone, K., Ram, N., & Bilalić, M. (2014). Playing off the curve-testing quantitative predictions of skill acquisition theories in development of chess performance. *Frontiers in psychology*, *5*, 923. doi:10.3389/fpsyg.2014.00923

Gray, W. D., & Lindstedt, J. K. (2017). Plateaus, dips, and leaps: Where to look for inventions and discoveries during skilled performance. *Cognitive science*, *41*(7), 1838-1870. doi:10.1111/cogs.12412

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493-498. doi:10.1111/2041-210X.12504

Haider, H., & Frensch, P. A. (2002). Why aggregated learning follows the power law of practice when individual learning does not: comment on Rickard (1997, 1999), Delaney et al.(1998), and Palmeri (1999). doi:10.1037/0278-7393.28.2.392

Harrell, F. E. (2015). *Regression Modeling Strategies*. Springer International Publishing.

Howard, R. W. (2014). Learning curves in highly skilled chess players: A test of the generality of the power law of practice. *Acta psychologica*, *151*, 16-23. doi:10.1016/j.actpsy.2014.05.013

Howard, R. W. (2018). Development of chess skill from domain entry to near asymptote. *The American Journal of Psychology*, *131*(3), 323-345. doi:10.5406/amerjpsyc.131.3.0323

Huang, J., Yan, E., Cheung, G., Nagappan, N., & Zimmermann, T. (2017). Master maker: Understanding gaming skill through practice and habit from gameplay behavior. *Topics in cognitive science*, *9*(2), 437-466. doi:10.1111/tops.12251

Keller, F. S. (1958). The phantom plateau. *Journal of the experimental analysis of behavior*, *1*(1), 1. https://doi.org/10.1901%2Fjeab.1958.1-1

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527. doi:10.1037/0033-295X.95.4.492

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & cognition*, *28*(5), 832-840. doi:10.3758/BF03198418

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition. *Cognitive skills and their acquisition*. (pp. 1- 55). Hillsdale, NJ: Erlbaum.

R Core Team (Version 4.2.0; 2022). R: A Language and Environment for Statistical Computing [Internet].

Rahman, R., & Gray, W. D. (2020). SpotLight on dynamics of individual learning. *Topics in Cognitive Science*, *12*(3), 975-991. doi:10.1111/tops.12512

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, *126*(3), 288. doi:10.1037/0096-3445.126.3.288

Robertson, R. J., & Glines, L. A. (1985). The phantom plateau returns. *Perceptual and motor skills*, *61*(1), 55-64. doi:10.2466/pms.1985.61.1.55

Saariluoma, P., & Hohlfeld, M. (1994). Apperception in chess players' long-range planning. *European Journal of Cognitive Psychology*, *6*(1), 1-22. doi:10.1080/09541449408520132

Thompson, J. J. C. (2018). *Use of digital records for studying skill learning* (Doctoral dissertation, Arts & Social Sciences: Department of Psychology). Simon Fraser University.

Thompson, J. J., Blair, M. R., Chen, L., & Henrey, A. J. (2013). Video game telemetry as a critical tool in the study of complex skill learning. *PloS one*, *8*(9), e75129. doi:10.1371/journal.pone.0075129

Thompson, J. J., Blair, M. R., & Henrey, A. J. (2014). Over the hill at 24: Persistent age-related cognitive-motor decline in reaction times in an ecologically valid video game task begins in early adulthood. *PloS one*, *9*(4), e94215. doi:10.1371/journal.pone.0094215

Thompson, J. J., McColeman, C. M., Blair, M. R., & Henrey, A. J. (2019). Classic motor chunking theory fails to account for behavioural diversity and speed in a complex naturalistic task. *PloS one*, *14*(6), e0218251. doi:10.1371/journal.pone.0218251

Van Der Maas, H. L., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *The American journal of psychology*, *118*(1), 29-60. doi:10.2307/30039042

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic bulletin & review*, *9*(2), 185-211. doi:10.3758/BF03196276

Appendix A.

Post Hoc Testing

Due to the importance of race in StarCraft 2, a weighted regression was run on the R^2 values for each model to determine if race influenced the fit of the models. In the base models, Zerg had significantly lower (β=-0.042, t=-2.067, p=0.042) R^2 values than Protoss and Terran. No race effects were detected for the models adding strategy. Terran had significantly higher (β=0.011, t=2.395, p=0.019) R^2 values than Protoss or Zerg in the models adding the interaction term.

Another weighted regression testing R^2 predicted by FAL was used to determine if skill affected the fit of the 3 models. No significant effects were detected.

All analyses were also tested using quadratic models (rather than the log-log method). The quadratic models improved the fit of the base model at the cost of needing to drop more games to keep reasonable power (following the recommendations of Harrell, 2015). There were similar insignificant observed effects for strategy and the interaction. No significant effects for FAL or race were detected.

Time to finish the third building was considered to capture the importance of timing in StarCraft 2 build orders. The time to finish the third building was added as a variable nested under categorical strategy. The R^2 was not significantly increased for either of the models of interest. There were no race or FAL effects detected in the models including time to finish third building.

Appendix B.

Countries Inhabited

Table B.1. List of countries inhabited by participants

Country Inhabited	Number of Participants
United States	21
Canada	7
United Kingdom	6
Germany	5
Poland	5
Finland	3
Netherlands	3
Norway	3
Australia	2
Czech Republic	2
Singapore	2
Sweden	2
Brazil	1
Chile	1
Denmark	1
France	1
Hungary	1
Israel	1
Italy	1
Mexico	1

Russian Federation	1
South Africa	1
Spain	1
Switzerland	1

Appendix C.

Model Outputs

Table C.1. Aggregated beta values for models with added strategy

Model Parameter	Weighted Mean Effect	Effect SE	n
Intercept: Barracks, Command Center, Barracks	2.894085348	0.004171712494	8
Intercept: Barracks, Command Center, Factory	2.761080557	0.005703784845	6
Intercept: Barracks, Factory, Starport	2.856177962	0.003349727868	6
Intercept: Forge, Nexus, Gateway	2.732914788	0.01181946779	1
Intercept: Gateway, Cybernetics Core, Gateway	2.969049277	0.002592085722	13
Intercept: Gateway, Cybernetics Core, Nexus	2.875240306	0.004039483721	7
Intercept: Gateway, Gateway, Cybernetics Core	3.114651238	0.01538813481	1
Intercept: Hatchery, Spawning Pool, Baneling Nest	2.889445323	0.009615524456	2
Intercept: Hatchery, Spawning Pool, Evolution Chamber	2.899109708	0.007609935757	3
Intercept: Hatchery, Spawning Pool, Hatchery	2.661998853	0.005883472331	4
Intercept: Hatchery, Spawning Pool, Roach Warren	2.763928405	0.006369770269	5
Intercept: Nexus, Forge, Gateway	3.041949045	0.0163538651	1
Intercept: Spawning Pool, Hatchery, Baneling Nest	2.78978736	0.005500638913	7
Intercept: Spawning Pool, Hatchery, Hatchery	2.674997025	0.00483719746	7
Intercept: Spawning Pool, Hatchery, Lair	3.115560485	0.0343586376	1
Intercept: Spawning Pool, Hatchery, Roach Warren	2.933816738	0.01122700218	2
Intercept: Spawning Pool, Roach Warren, Hatchery	2.971434516	0.03110329559	1
Intercept: Spawning Pool, Roach Warren, Lair	2.88407429	0.03605492453	1
LOGdays	-0.05079902658	0.0006056805113	76
Barracks, Command Center, Barracks	-0.01070402458	0.002354494062	10
Barracks, Command Center, Factory	-0.006037721069	0.001105708301	14
Barracks, Factory, Command Center	-0.01762250267	0.003258649455	2
Barracks, Factory, Starport	0.01192354324	0.001517729641	8
Command Center, Barracks, Factory	-0.02045539661	0.00489183434	2
Gateway, Cybernetics Core, Gateway	0.01104257347	0.001430433461	10
Gateway, Cybernetics Core, Nexus	-0.01064628196	0.001153828498	16

Gateway, Cybernetics Core, Robotics Facility	0.006197975461	0.002455751775	6
Gateway, Cybernetics Core, Stargate	0.0038585181	0.004926872801	1
Gateway, Cybernetics Core, Twilight Council	0.001679300891	0.004909981413	2
Gateway, Forge, Cybernetics Core	-0.01552223571	0.007894720684	1
Gateway, Gateway, Cybernetics Core	-0.003050051136	0.007370381354	1
Hatchery, Spawning Pool, Evolution Chamber	-0.01568523246	0.008563238073	1
Hatchery, Spawning Pool, Hatchery	-0.001185356607	0.002089112352	5
Hatchery, Spawning Pool, Lair	-0.001573078792	0.003378586005	4
Hatchery, Spawning Pool, Roach Warren	-0.0132091495	0.003392862202	5
Nexus, Forge, Gateway	-0.002194637785	0.001895931433	7
Spawning Pool, Baneling Nest, Hatchery	0.00728756954	0.005546898469	1
Spawning Pool, Hatchery, Baneling Nest	0.02371424982	0.002327144152	8
Spawning Pool, Hatchery, Evolution Chamber	-0.0007634879651	0.006330383092	2
Spawning Pool, Hatchery, Hatchery	0.0003731384169	0.001263613398	22
Spawning Pool, Hatchery, Lair	-0.01581188424	0.005472934601	2
Spawning Pool, Hatchery, Roach Warren	-0.0006212158416	0.002332354273	8
Spawning Pool, Roach Warren, Hatchery	0.02718130471	0.01023355077	2
Spawning Pool, Roach Warren, Lair	0.02510723591	0.01018240473	2

Table C.2. Aggregated beta values for models with added interaction

Model Parameter	Weighted Mean Effect	Effect SE	n
Intercept: Barracks, Command Center, Barracks	2.882756015	0.005452978671	8
Intercept: Barracks, Command Center, Factory	2.762143179	0.01102054393	6
Intercept: Barracks, Factory, Starport	2.848083263	0.005362629281	6
Intercept: Forge, Nexus, Gateway	2.735863562	0.0179112216	1
Intercept: Gateway, Cybernetics Core, Gateway	2.929833867	0.005231218204	13
Intercept: Gateway, Cybernetics Core, Nexus	2.901531987	0.006141954205	7
Intercept: Gateway, Gateway, Cybernetics Core	3.104837114	0.01943875605	1
Intercept: Hatchery, Spawning Pool, Baneling Nest	2.912956603	0.01527827882	2
Intercept: Hatchery, Spawning Pool, Evolution Chamber	2.918213311	0.01091322349	3
Intercept: Hatchery, Spawning Pool, Hatchery	2.650213971	0.007939086443	4
Intercept: Hatchery, Spawning Pool, Roach Warren	2.802814337	0.01069452826	5

Intercept: Nexus, Forge, Gateway	3.075980822	0.02772062426	1
Intercept: Spawning Pool, Hatchery, Baneling Nest	2.785932117	0.008250623215	7
Intercept: Spawning Pool, Hatchery, Hatchery	2.689587097	0.008919889664	7
Intercept: Spawning Pool, Hatchery, Lair	3.152290898	0.05173420568	1
Intercept: Spawning Pool, Hatchery, Roach Warren	3.056841598	0.03441040797	2
Intercept: Spawning Pool, Roach Warren, Hatchery	3.168933891	0.07283618129	1
Intercept: Spawning Pool, Roach Warren, Lair	2.352473074	0.2455269314	1
LOGdays	-0.04553389185	0.001076976484	76
Barracks, Command Center, Barracks	0.03390526979	0.01214336295	10
Barracks, Command Center, Factory	0.04779830962	0.007279972871	14
Barracks, Factory, Command Center	0.001305377729	0.02528618668	2
Barracks, Factory, Starport	-0.006992471806	0.007045719333	8
Command Center, Barracks, Factory	-0.02805265357	0.02055336506	2
Gateway, Cybernetics Core, Gateway	-0.009296178746	0.01166128849	10
Gateway, Cybernetics Core, Nexus	-0.01729653665	0.006297100695	16
Gateway, Cybernetics Core, Robotics Facility	0.02765053619	0.02107682771	6
Gateway, Cybernetics Core, Stargate	-0.01270316694	0.02114155433	1
Gateway, Cybernetics Core, Twilight Council	0.02680133417	0.02493238211	2
Gateway, Forge, Cybernetics Core	-0.06378160694	0.03251880933	1
Gateway, Gateway, Cybernetics Core	-0.01423289636	0.04350484372	1
Hatchery, Spawning Pool, Evolution Chamber	-0.121466232	0.05102401138	1
Hatchery, Spawning Pool, Hatchery	-0.0100155046	0.01422248717	5
Hatchery, Spawning Pool, Lair	-0.02844992908	0.01696573638	4
Hatchery, Spawning Pool, Roach Warren	-0.02580188628	0.0172525508	5
Nexus, Forge, Gateway	-0.02008273399	0.01213180503	7
Spawning Pool, Baneling Nest, Hatchery	-0.05011265949	0.05307774478	1
Spawning Pool, Hatchery, Baneling Nest	0.02528054993	0.01100924361	8
Spawning Pool, Hatchery, Evolution Chamber	-0.01245760286	0.02190853344	2
Spawning Pool, Hatchery, Hatchery	-0.01003916735	0.007184664953	22
Spawning Pool, Hatchery, Lair	0.00357107113	0.0261178112	2
Spawning Pool, Hatchery, Roach Warren	-0.04964878427	0.01093832952	8
Spawning Pool, Roach Warren, Hatchery	-0.1282632897	0.05404817506	2
Spawning Pool, Roach Warren, Lair	0.05632380054	0.07947349565	2

LOGdays:Barracks, Command Center, Barracks	-0.02982050802	0.006626856207	10
LOGdays:Barracks, Command Center, Factory	-0.02803895645	0.003474222974	14
LOGdays:Barracks, Factory, Command Center	-0.01459648962	0.01244575308	2
LOGdays:Barracks, Factory, Starport	0.008452244002	0.003345069779	8
LOGdays:Command Center, Barracks, Factory	-0.003092017202	0.01704127067	2
LOGdays:Gateway, Cybernetics Core, Gateway	0.01252447767	0.005449149375	10
LOGdays:Gateway, Cybernetics Core, Nexus	0.004382340619	0.00319690807	16
LOGdays:Gateway, Cybernetics Core, Robotics Facility	-0.01338740282	0.009312326439	6
LOGdays:Gateway, Cybernetics Core, Stargate	0.008755944165	0.01165852235	1
LOGdays:Gateway, Cybernetics Core, Twilight Council	-0.0107178995	0.01204594088	2
LOGdays:Gateway, Forge, Cybernetics Core	0.02381365195	0.01443842952	1
LOGdays:Gateway, Gateway, Cybernetics Core	0.005281509564	0.01950649179	1
LOGdays:Hatchery, Spawning Pool, Evolution Chamber	0.05392666321	0.02576456079	1
LOGdays:Hatchery, Spawning Pool, Hatchery	0.001749956758	0.006707397895	5
LOGdays:Hatchery, Spawning Pool, Lair	0.01230978935	0.009418189986	4
LOGdays:Hatchery, Spawning Pool, Roach Warren	0.01197117601	0.009533706423	5
LOGdays:Nexus, Forge, Gateway	0.0120794721	0.005944080906	7
LOGdays:Spawning Pool, Baneling Nest, Hatchery	0.02320755281	0.021438015	1
LOGdays:Spawning Pool, Hatchery, Baneling Nest	-0.006769233934	0.005514941308	8
LOGdays:Spawning Pool, Hatchery, Evolution Chamber	0.01353398204	0.01293738549	2
LOGdays:Spawning Pool, Hatchery, Hatchery	0.005090849312	0.003712101815	22
LOGdays:Spawning Pool, Hatchery, Lair	-0.001922607111	0.01295524113	2
LOGdays:Spawning Pool, Hatchery, Roach Warren	0.02718550005	0.005329120634	8
LOGdays:Spawning Pool, Roach Warren, Hatchery	0.08064887032	0.02388395186	2
LOGdays:Spawning Pool, Roach Warren, Lair	-0.03353412736	0.04515488995	2