Use of Digital Records for Studying Skill Learning

by

Joseph J Thompson

M.A., Simon Fraser University, 2011
B.A., Simon Fraser University, 2004

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Psychology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY
Spring 2018

Approval

Name:	**Joseph J Thompson**
Degree:	**Doctor of Philosophy (Psychology)**
Title:	**Use of Digital Records for Studying Skill Learning**
Examining Committee:	**Chair:** Thomas Spalek Professor
	Mark Blair Senior Supervisor Associate Professor
	Timothy Racine Supervisor Professor
	Kathleen Slaney Supervisor Associate Professor
	Maite Taboada Internal Examiner Professor Department of Linguistics
	Tom Stafford External Examiner Senior Lecturer Department of Psychology University of Sheffield
Date Defended:	**April 27, 2018**

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.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

The present work uses a novel data source, real-time strategy video game play in StarCraft 2, to study complex skill learning. Chapter One discusses some important desiderata of a large dataset. Chapter Two discusses domain specifics about StarCraft 2, and introduces the process by which survey respondents donate digital archives which are parsed to reveal second-by-second information about in-game performance of players. Chapter Three asks how experience should be defined in a complex domain. I find that the common-sense definition, that experience should be measured soley in terms ot task-specific experience, misleads researchers by being both overly permissive and restrictive. A better definition can be achieved by focusing on other forms of experience, such as experience with different game modes. Chapter Four extends a previous study of age-related declines in a StarCraft 2 cross-sectional dataset. Segmented regression models are used to estimate the onset of age-related differences. Secondly, I examine the theory that large swaths of age-related differences, across a wide array of variables, are attributable to a single general cognitive, but not psychomotor, factor. I find support for this theory, as a simplified measure of redundant click-speed accounts for about 19% of the shared age-related variance in established measures of StarCraft 2 speed. In Chapter Five I examine some of the common responses to the idea that Big Data, and the emerging data sources they employ, could effectively replace the role of theory in science. I argue, instead, that emerging data sources are a threat to overzealous generalizations from laboratory grown theories to complex behaviour. If emerging data sources fulfill their potential as tools for evaluating theory generality, then scientific standards for making claims about generality could change in pronounced ways. This would create a bigger gap between empirically grounded generalizations from the laboratory to life and careless generalizations which Frankfurt would call "bullshit." Finally, I examine two very different research strategies for going about the evaluation of theory using Big Data, and point to the virtues and limitations of both.

Keywords: Expertise; Naturalistic Telemetry; Video Games; Skill learning; Aging

Acknowledgements

I would like to thank my entire Committee for their useful feedback and guidance over the course of my graduate education. Kathleen Slaney has been helpful in showing me how to navigate issues in theoretical psychology. Special thanks goes to Timothy Racine not only for continually challenging my assumptions about psychology and philosophy, but also for his continued mentorship since I expressed interest in specializing in psychology. Of course, I also owe the deepest gratitude to Mark Blair for the tremendous support he has provided me over the years.

The support of the Cognitive Science Laboratory at SFU was also integral to this dissertation, especially the help from Cal Woodruff in processing the longitudinal data. Thanks also goes to Caitlyn McColeman, Jordan Barnes, Katerina Stepanova, Kat Dolguikh, Nathan Hutchinson, Robin Barrett, Yue Chen, David McIntyre, Alex Volkanov, Scott Harrison, Judi Azmand, Neda Anvari, Romanos Byliris, Ruilin Zhang, Calvin Chou, Rajan Hayre, Judi Azmand, Ruth Jen, and all the members of the cognitive science lab. Special thanks also goes to Nehdia Sameen, Donna Tafreshi, and Tyler Wereha for their influence on my thinking.

Importantly, I want to thank an anonymous commentator on my poster at the 2014 American Psychological Society conference in San Francisco, who suggested that Salthouse might be interested in Finger Tapping data in StarCraft 2. Years later this lead me to the papers from Salthouse which inspired Chapter Four. Finally, I want to thank the Social Sciences and Humanities Research Council of Canada for their support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the Big Data revolution there have been a wide array of datasets made available to researchers, many of which are particularly valuable for the evaluation of psychological theory. These are sometimes called naturally occurring datasets (Goldstone & Lupyan, 2016) or, where the researcher is employing data which is automatically recorded by the machines used by participants outside of the laboratory, naturalistic telemetry (Thompson et al., 2013). The present work will discuss two independent empirical projects utilizing these data sources to study skill learning. It will also address the broader question of whether the Big Data revolution is a boon for psychology. However, given that the use of these data sources is still relatively new to psychology (Goldstone & Lupyan, 2016; Markowetz, Błaszkiewicz, Montag, Switala, & Schlaepfer, 2014; Thompson et al., 2013), I postpone the question of Big Data's relevance to psychology until after a presentation of the empirical work.

While the present work begins with two empirical studies before turning to broader theoretical and methodological issues, a brief discussion of the importance of theoretical considerations in the evaluation of method is necessary. This is largely because there have been provocative claims about the capacity of Big Data to replace theory (C. Anderson, 2008). I examine these provocative claims in more detail in Chapter Five but, for the present purposes, I begin with a discussion of three key desiderata of a naturally occurring dataset:

1. Size
2. Indicators of sample representativeness
3. Longitudinal data

Rather than propose a more exhaustive list, as has been attempted elsewhere (Goldstone & Lupyan, 2016), I focus on the importance of theory at the level of method. I will then outline the specific objectives of the present work, taking for granted that even the exploratory aspects of the present work is at least somewhat theory-laden.

1.0.1 Size

Large samples are usually an advantage regardless of whether we are thinking about dataset size in terms of the number of observations per variable (i.e., rows) or the number of variables (i.e., columns). While there are cases where sample sizes cause methodological challenges, these challenges can be addressed. For example, one common concern is that Big Data has a capacity to detect extremely small effects, which means that many non-directional hypothesis tests will be statistically significant even if the effect sizes are too small to be important (Lin & Lucas, 2013). Even here, one is free to report effect sizes or, if appropriate, choose a hypothesis test for which the detection of small effects is actually of interest. However, data collection can be expensive, and consequently theoretical motivations quickly become relevant to whether dataset size is worth pursuing. This is most obvious in variable selection, where we have much more patience for the collection of data on covariates of direct theoretical relevance. Patience quickly wanes when collecting extraneous variables which can be ignored without any obvious repercussion, even if they are potentially useful for followup exploratory analyses. Consequently, size is usually advantageous but expensive, so researchers must draw on theory to determine whether this investment is worthwhile.

1.0.2 Indicators of Sample representativeness

Sampling bias can be troublesome in telemetry because (a) researchers do not collect data in their own labs using their own machines and (b) the Application-Programming-Interfaces (APIs) made available by companies which provide access to data may not be understood by the researcher. Some APIs used for collecting Twitter data, for example, leave large amounts of data out and, more importantly, the basis of exclusion may not be random (Boyd & Crawford, 2012). Furthermore, datasets scraped off the web will lack data on basic demographics, making it even more difficult to understand whether sampling has been biased.

If one is willing to put in additional effort, one can secure variables that can help ascertain whether sampling was biased. The dataset from Thompson et al. (2013), and indeed all data presented in this work, are based on a hybrid of online surveys and telemetry. This allows participants to voluntarily donate their data while also allowing researchers the opportunity to collect demographic information (see Chapter Two for details). This information can be used to ask whether the sample is representative (e.g., it may be that certain genders were more likely to go to the website where the survey was posted). The source of sampling bias in these hybrid methods are of the more traditional sort. The participants sampled are those that (a) researchers can reach, and in this case reach online, and which (b) are willing to participate.

What is most important for the present purposes is that the choice of method will once again be shaped by theoretical considerations. The preference for random sampling

might appear theory-neutral insofar as it follows from basic research methodology and its underlying philosophy of science. However, random sampling is often impossible to achieve in practice. In these cases, researchers also need to weigh costs and benefits of collecting data through different methods. I distinguish four for the present purposes:

1. Traditional sampling methods (e.g., advertised surveys)
2. Data downloaded using tools provided by companies and third parties
3. Scraped data (e.g., using a program to extract web data)
4. Hybrid approaches

True random sampling may or may no be possible in the above cases. Identifying the method which produces the least vicious bias is, at least in part, theory-laden. I will use the example of StarCraft 2 throughout the discussion, but the lesson would hold for any domain.

One important theoretical decision is whether the data-collection tools offered by companies will lead to greater or lesser sampling bias than the use of online surveys. Online surveys which advertise to participant video gamers are likely to have more highly committed respondents (Khazaal et al., 2014), which may explain the large number of skilled individuals in earlier datasets on skilled StarCraft 2 play (Thompson et al., 2013). Another option would be to 'scrape' digital StarCraft 2 records off the web using a program designed to find and download these files. Scraping has two disadvantages in the case of studying StarCraft 2, however. First, theory might suggest that scraped sampling is likely to be even more biased in the context of StarCraft 2 research, as this would only sample individuals willing to post their records publicly. More serious players are presumably more likely to upload only their best performance. Secondly, the lack of demographic data makes the understanding of sample bias more difficult. For example, while it appears that women make up a substantial proportion of video games (*Essential facts about the computer and video game industry*, 2017; *Video games in europe: Consumer study*, 2012), Thompson et al. (2013) were nevertheless surprised to find that their sample of 3,395 StarCraft 2 players only contained 29 women, and this knowledge allowed them to restrict any generalizations to male participants. Scraping methods would have left them blind to these sorts of obstacles to generalization.

In other cases, there may be clear advantages to relying on industry data sources. For example, Whitehead and Perry (2017) wanted to understand how pornography consumption varied by religious background, a question that may be difficult to address using traditional methods given how the impact of social desirability bias is likely to vary with religious factors. They used aggregated data on pornography searches by state, which were made available by Google, to sidestep the plausible influence of social desirability bias on responses.

1.0.3 Longitudinal data

Naturalistic telemetry allows for the collection of longitudinal datasets that are impossible to acquire by other means, and the emergence of these data is likely to be more important for some purposes and less important for others. In the domain of skill research, theoretical issues will likely determine the importance of longitudinal data. For example, classic research on expertise is dominated by expert-novice comparisons which are the most insightful when expertise is characterized as the simultaneous automatization of subskills. If subskills develop simultaneously and develop at similar rates, then we can simply look at the end points of skill learning and reasonably interpolate a developmental story. But it is not clear that expertise behaves this way outside of very simple tasks. Experts differ from novices in the provision of educational and cultural resources (Ericsson, Krampe, & Tesch-roemer, 1993). Consequently, to look at cross-sectional comparisons of experts and novices, while certainly insightful for certain purposes, is to sample time slices from qualitatively different developmental stories. From a different theoretical perspective, interpolation of learning trajectories would be much less contentious.

1.0.4 Goals of the present work

While the abundance of naturalistic telemetry brings new opportunities for science, it has only been utilized in a handful of skill learning studies (Huang, Yan, Cheung, Nagappan, & Zimmermann, 2017; Lewis, Trinh, & Kirsh, 2011; Stafford & Haasnoot, 2017; Thompson et al., 2013, 2014; Thompson, McColeman, Stepanova, & Blair, 2017). The present study aims to extend prior work by using datasets that have the previously mentioned desiderata. In Chapter Two, I will describe the methods used to collect the required data. All the studies presented here will enjoy large sample sizes and at least some access to basic demographic information. In Chapter Three, I will utilize longitudinal data in studying expertise and, in Chapter Four, I will employ cross-sectional data which are particularly valuable for testing the generalizability of theories about age-related change. In Chapter Five I will return to the broader methodological issue of Big Data's value to the cognitive sciences of skill learning.

What is experience?

Chapter Three deals with the question of how one might define experience in a longitudinal study of skill. A brief review of the literature will reveal that this question is important and that its answer is not necessarily trivial. This is best seen by looking at an example of a study using naturalistic telemetry to address foundational questions in the psychology of skilled performance.

Longitudinal studies of skill learning are, of course, not new (Ericsson et al., 1993; Thorndike, 1908). However, the application of Big Data to the study of skill learning is still in its very early stages. One attempt has been made by Huang et al. (2017), who had

access to 3.2 million matches of the first-person shooter action video game *Halo Reach*. More impressive still, the dataset was a complete census of this game's first seven months. Such data are of clear theoretical value to psychology. For example, the nature of discontinuities in skill learning is a literature spanning 120 years (Bryan & Harter, 1897, 1899; Gray & Lindstedt, 2016; Keller, 1958). Bryan and Harter (1897, 1899) described discontinuities in telegraphy skill where listeners struggled to decode incoming messages as fast as they could be received.

> For many weeks there is an improvement which the student can feel sure of and which is proved by objective tests. Then follows a long period when the student can feel no improvement, and when objective tests show little or none. At the last end of the plateau the messages on the main line are, according to the unanimous testimony of all who have experience in the matter, a senseless clatter to the student-practically as unintelligible as the same messages were months before. Suddenly, within a few days, the change comes, and the sense-less clatter becomes intelligible speech (Bryan & Harter, 1897, p. 52-53).

Such discontinuities in skill learning, if they exist, have deep theoretical significance. They are predicted, for example, by the view that expertise requires a hierarchy of habits whereby lower-level skills must be mastered before moving onto higher-level skills.

> A hypothesis set forth to explain the plateau in learning is that there is a hierarchy of habits to be mastered by the individual when he attempts to learn a complex task. After succeeding in the first order, he may be fixated at that level for some time before becoming able to integrate the patterns needed in the second-order habits. Information is consolidated and reorganized. An example of this situation is found in tennis. First-order habits such as learning to stroke the ball when in a stationary position. Second-order habits could include hitting the ball while on the move, and a third-order category might include the integration of effective movement patterns in the game situation. Theoretically, depending on the manner in which the sport is taught and the performer involved, a plateau could occur at any one of these transitional periods (Singer, 1975, p.129-130).

In other words, a hierarchy of habits could produce discontinuities in skill learning curves because, once the prerequisite skill in lower-level habits are reached, or once higher level skills start being trained, new learning effects associated with higher-level habits come into being. The existence and timing of plateaus might then provide crucial clues as to how researchers understand complex tasks. Positing of transitional periods or low-level subskills should be accompanied by the presence of plateaus (at least for some individuals). Unfortunately it is difficult to acquire datasets which confirm their existence, but see Gray and Lindstedt

(2016) and Donner and Hardy (2015) for a discussion of some examples. Crucially, Huang et al. (2017) do find plateaus for some individuals and not others.

Ideally, future work can focus in on frequently occurring plateaus and clarify their nature. Gray and Lindstedt (2016), for example, argue that many peaks in performance actually reflect the limitations of a method for accomplishing the task. These have been called 'spurious plateaus' (Gray & Lindstedt, 2016; Thorndike, 1913). Consequently, probing plateaus (or transient dips in performance) could provide hints as to the methods being used by participants, the motivational factors which allow individuals to practice through difficult times, and perhaps even the process by which individuals discover entirely new methods of accomplishing a task.

It is clear that the research of Huang et al. (2017) is of great potential value to psychology, but it is important to note that this work requires assumptions about how to define experience in *Halo Reach*. Huang and colleagues do not simply define experience as the number of Halo Reach matches played. They recognize that there are multiple game modes in *Halo Reach* such as 'Slayer' and 'capture the flag,' each with different objectives. The authors feared that skill might not transfer from one game mode to another, so they restricted their analysis to 'Slayer' as this was the most common game mode. This choice betrays an implicit assumption that there will be little or no transfer between game modes.

The implicit assumption that expertise exhibits *extreme domain-specificity*, by which I mean that skill in one domain will not transfer to others, is entirely warranted. Indeed, extreme domain-specificity is considered common wisdom in expertise research. As Feltovich, Prietula, and Ericsson (2006) point out, the oldest empirical evidence of the extreme domain-specificity of skill is probably Ebbinghaus (1885), who found substantial differences in memory performance when using nonsense words. Taking individuals out of their domain of proficiency (i.e., language) eliminated a domain-specific memory advantage. Today expertise is seen as an optimization of the cognitive system to the task at hand (Ericsson & Lehmann, 1996), so one should not assume transfer when task demands change even slightly. Keetch, Lee, and Schmidt (2008), for example, have found that shots from the free-throw line in basketball are more accurate than one would predict from the accuracy of closer and farther shots. Basketball players have spent so much time practicing this particular kind of shot that they have developed a foul-line shot expertise which is embedded in the more general domain of basketball (Keetch et al., 2008). This embedded experience clearly does not perfectly transfer to shots from other distances or angles, or even to jump shots from the same location (Keetch et al., 2008). If I can find failures of transfer between similar behaviours (i.e., basketball throws) within the same domain, then it makes sense to be skeptical that skill will transfer between loosely similar modes of video game play.

Huang et al. (2017), therefore made a sensible decision to restrict their analysis to one game mode within *Halo Reach*. What remains problematic, however, is that skills do sometimes transfer, often in ways that are difficult to foresee. Basketball players exhibit

better than average accuracy in dart throwing, but lack attentional indicators of expertise found in skilled basketball and skilled dart throwing performance (Rienhoff et al., 2013). An even more dramatic example comes from the literature growing around the question of whether video game practice transfers to basic cognitive abilities (Green & Bavelier, 2003; Green et al., 2017; Redick, Unsworth, Kane, & Hambrick, 2017; Unsworth et al., 2015). A review of the literature about the relationship between basic cognitive abilities and video game performance is beyond my present scope. However, the fact that these transfer effects are still debated after almost fifteen years of research is further evidence that transfer-effects can be counter-intuitive, and therefore questions of transfer should be resolved empirically on a case-by-case basis.

A major challenge that the existence of transfer brings to expertise research is that otherwise sensible definitions of experience may be inadequate in light of unintuitive facts about skill transfer. Huang et al. (2017) have good reason to think that different game modes of Halo Reach will not transfer perfectly as the players are likely optimizing their cognitive systems to entirely different objectives. But it remains possible that there are important sources of imperfect transfer across game modes. More problematic still is the possibility that some players have different play-styles, such as 'run around shooting everything that moves'. Some play-styles might transfer across game modes better than others.

Consequently, future work using Big Longitudinal Data will undoubtedly rely on increasingly careful operationalizations of experience, as how experience is defined obviously impacts the learning curves which are typically of interest. In the case of examining plateaus of skill learning, including irrelevant experience in the operationalization could produce discontinuities in learning curves. Players who intersperse game modes ignored by the researcher into their play, and learn from this experience, will exhibit transient speedups in learning. Researchers might incorrectly assume that these speedups reflect moving beyond a 'spurious plateau' (Gray & Lindstedt, 2016; Thorndike, 1913), or a limitation on performance that individuals can surpass with more motivation and effort or a better method. Since examining the learning curves of skill acquisition assumes an adequate operationalization of experience, and since the shape of learning curves has been such a prolific area of research, it makes sense that early longitudinal explorations of naturalistic telemetry should focus explicitly on the question of how experience should be defined in the context of StarCraft 2 research on skill. Chapter Three will be focused on selecting an empirically informed operationalization of experience and reporting some preliminary findings.

Finger tapping in StarCraft 2

While there is a prima facie case to be made that naturalistic telemetry can be a useful tool for examining the generalizability of psychological theory, there are still only a handful of examples. One useful example is J. R. Anderson and Schooler (1991) research using New York Times headlines, existing databases of children and parents talking, and a collection of

emails. They showed that classic effects of frequency, recency, and spacing which typically explain memory performance (Ebbinghaus, 1885) also predict the probability that a word will appear on the 101st day of study. This invites the speculation that our memory system has the learning and retention curves that it does because our memory system is adapted to remember items that we are most likely to need. Rather than reflexively viewing the properties of forgetfulness as a limitation of the cognitive system, such work could help reorient our thinking about findings in the memory literature. This work could contribute to theory in a number of indirect ways, but one specific impact is that it could alleviate the need for a theory of why the memory system has the limitations it does. Chapter Four attempts to provide a similar contribution to the study of age-related change using naturalistic telemetry from StarCraft 2 data.

Not surprisingly, some theories will generalize better than others when examined from the purview of StarCraft 2 data. For example, Thompson et al. (2014) examined whether the cross-sectional findings of age-related decline (Salthouse, 2009; Schroeder & Salthouse, 2004; Tsang & Shaner, 1998) would be found in the context of StarCraft 2 response speeds. Some of the predicted effects, such as the predicted slowings in response times across age, were found. Other expected effects, such as age-related slowing in dual-task performance (Verhaeghen, Steitz, Sliwinski, & Cerella, 2003), or the capacity of expertise to prevent age-related slowing in response speeds (Bosman, 1993; Salthouse, 1984), were not found. The failure of expertise to ameliorate slowing was especially surprising as older typists appear to overcome the effect of age by reading farther ahead (Bosman, 1993), and StarCraft 2 contains a rich interface that presumably provides many more avenues of compensation than a type-writer. The failure to find declines in dual-task performance was similarly puzzling. One possibility was that some unknown factor was bolstering the dual-task performance of older individuals, obfuscating declines.[1]

Chapter Four extends the previous analysis of age-related differences in StarCraft 2 (Thompson et al., 2014). While the previous work presented compelling evidence that StarCraft 2 players slow with age, this finding had relatively narrow theoretical relevance. It is useful to know that the cross-sectional laboratory declines (Salthouse, 2009; Schroeder & Salthouse, 2004; Tsang & Shaner, 1998) occurring in people's 20s may be impacting performance in much more complex task domains, but these results do not speak in favour of any particular *explanation* for these declines. Chapter Four targets the processing speed theory from Salthouse (1996), which implies that many of the observed age-related declines are due to a single underlying construct that Salthouse called 'processing speed.' I develop a speed measure in StarCraft 2, Right-Click Latency, which is potentially influenced by processing speed yet presumably imposes far fewer cognitive demands on the participant. I then probe

[1]Of course, it is logically possible that our best theories of aging simply do not generalize as well as we had hoped.

it for possible age-related decline, and consider whether it might explain the relationships observed in Thompson et al. (2014).

Chapter 2

Methods

2.1 StarCraft 2

StarCraft 2 is a real-time strategy game where players manage a civilization and try to destroy an opponent's civilization. This requires balancing the needs of collecting resources with the creation and maintenance of an army. There are a few features of the game which are especially relevant to cognitive scientists. While I postpone the details of how StarCraft 2 data are collected until later in this chapter, readers can assume that comprehensive data are available at the level of second-by-second performance (one observation per action) and at the level of the game-level (one observation per game).

2.1.1 Features of StarCraft 2

1. StarCraft 2 occurs in real time (making speed advantageous).

2. Most army commands in StarCraft 2 are reversible (the accuracy of clicks is very important only rarely).

3. Games begin with the same starting conditions for all players, which includes a base (Figure 2.1 D) and a handful of workers (Figure 2.1 E).

4. To command pieces, players must *select* the units with their mouse. The player in Figure 2.1 has selected the unit at location E, and the game confirms this selection in panel B. The available commands for this unit are displayed in Figure 2.1 C. Managing units becomes more demanding as civilizations grow, and strong players can assign units to hotkeys which allows them to select these units with a keystroke.

5. Balancing military and economic considerations requires frequent task switching (three kinds of task involve ordering the production of workers (Figure 2.1 E), expansion of the civilization to new bases (Figure 2.1 D), and movement of the army (Figure 2.1 F).

Figure 2.1: An example StarCraft 2 play as the 'Protoss' race. A: a miniature map that contains gross information about the entire play area. Gross information becomes available as players move their units across the map. B: information about the currently selected unit (a probe). C: possible commands for the selected unit (Commands can be issued by mouse or by keystroke). D: the command structure for Protoss. This command structure can be used to produce Probes. E: a selected probe, which is used to collect resources and operationalization new buildings. F: a military unit.

6. Players only receive gross information about the current game-state from the minimap (Figure 2.1 A), and must therefore shift the *viewscreen* (the area occupied by Figure 2.1 D,E, and F) to allocate their attention to a new portion of the game map. The interface allows for more and less efficient ways of moving the view screen. Methods include:

   (a) placing the mouse in the corner of the viewscreen to drag the screen gradually.

   (b) clicking on a portion of the minimap with the mouse to place the viewscreen over that location.

   (c) double-tapping a hotkey associated with a unit to place the screen over top of the unit.

7. Regardless of how a player moves their screen, they have imperfect information about the game-state (see black area in Figure 2.1 A). Only when units move around the map can players receive detailed information about their opponent's actions.

8. StarCraft 2 players can play as one of three unique 'races', which differ slightly in terms of game-play mechanics.

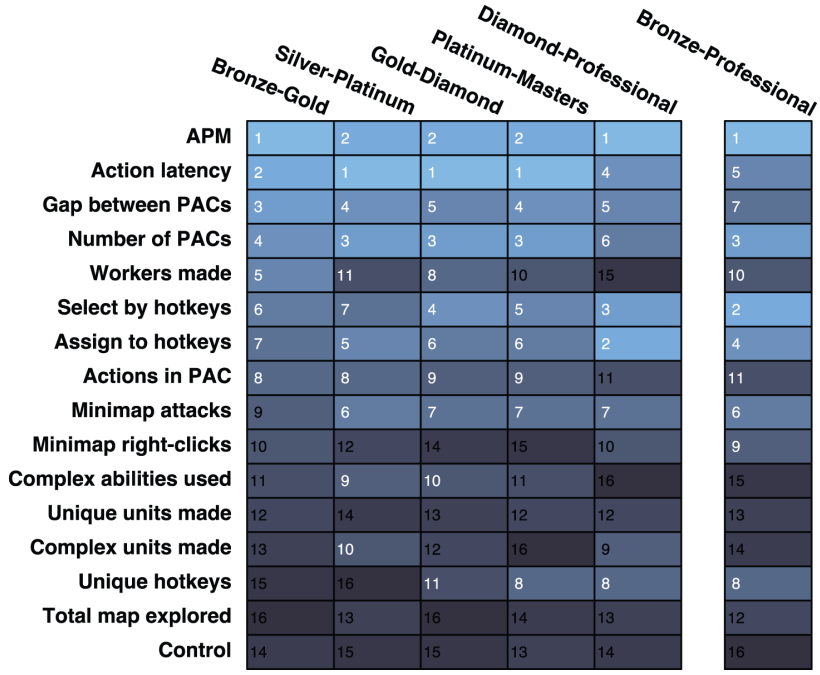


Figure 2.2: A depiction of predictive importance of fifteen performance variables from Thompson et al. (2013). Skill is defined in terms of eight increasingly prestigious leagues ('Bronze', 'Silver', 'Gold', 'Platinum, 'Diamond', 'Master', 'Grandmaster', and 'Professional'). For more information on skill, see Section 2.2. Each column depicts 25 conditional inference forest binary classifications of skill using fifteen performance variables and a random noise control. Within-cell numbers reflect the ranking predictive power, based on the median permutation importance index of 25 runs of the classifier. White numbers depict significantly predictive variables based on the criteria found in Linkletter et al. (2006). Variables are significantly predictive iff their median permutation importance score is above the 95th percentile of the control.

Even this abbreviated description of StarCraft 2 reveals much of psychological interest in the game. Features 1 and 2 suggest that StarCraft 2 is a good domain for testing cognitive-motor speed as the importance of accuracy is relatively less. This makes speed-accuracy trade-offs less likely. Feature 4 implies that players must make good use of the game interface to manage the complexity of a larger civilization. Indeed, Thompson et al. (2013) found that the use of hotkeys to select units was predictive of skill in StarCraft 2 (also see Figure 2.2). Feature 5 betrays the dual-task nature of the game, and this feature colours the interpretation of any StarCraft 2 meta-data. For example, Thompson et al. (2013) found

that the number of workers produced was an important predictor of StarCraft 2 skill when distinguishing novices. They hypothesized that these differences in the number of workers trained was due to overwhelmed cognitive load (Thompson et al., 2014). Weaker players were unable to keep up the periodic production of workers without being distracted by other aspects of the game.

Features 6 and 7 are foundational for the present work. Screen movements in StarCraft 2 can be very chaotic (especially when the screen is moved using Feature 6a). Consequently, screen movement data in StarCraft 2 present some of the same methodological problems as raw gaze data in eye-tracking research (see Figure 2.3). To deal with this problem I use an dispersion-threshold algorithm (Salvucci & Goldberg, 2000) which aggregates raw attentional data into *screen fixations*, periods of time where player screens are falling within a relatively small area. The algorithm proceeds by taking tentatively placing screen movements within a fixation and calculating the smallest rectangle which would encompass all of the screen movements in the fixation. If the sum of the length and width of the rectangle is greater than 6 game coordinates, then the screen movement is removed from the list of movements associated with that fixation. Finally, if the final collection of screen movements occupies more than twenty game timestamps (about 226 milliseconds), I posit a fixation. The fixations duration is determined by the time span taken up by the component screen movements, and its location is determined by averaging the coordinates of the screen movements.

The definition of fixations is foundational to the present work. The time from the onset of a fixation to the first action within the fixation (assuming there is such an action) is called the screen fixations *First Action Latency* (or FAL). There are many fixations per game, and aggregations of First Action Latencies have been used to predict skill (Thompson et al., 2013), examine age-related differences (Thompson et al., 2014), and examine whether motor control theories generalize well to StarCraft 2 (Thompson, McColeman, et al., 2017). However, while previous work has been content to use mean First Action Latencies, the present work will rely on *median FALs*, as within-game distributions of First Action Latencies are usually strongly skewed.

Feature 8 has been a nuisance variable in previous work on StarCraft 2 (Thompson, McColeman, et al., 2017), as researchers have been interested in more general claims about StarCraft 2 performance, rather than subtle race differences. In StarCraft 2, players must play as one of three different 'races,' each with subtle gameplay differences. For example, the 'Terran' pieces are trained from flying production structures, the 'Zerg' pieces are grown from larva that appear at the command structure, and a portion of the 'Protoss' pieces can emerge from any location the player chooses (as long as this location is close to certain Protoss buildings). The races also have different units with different strengths and weaknesses. These differences appear to come with some performance differences, with Zerg players typically appearing faster. On the other hand, all races have command structures which produce

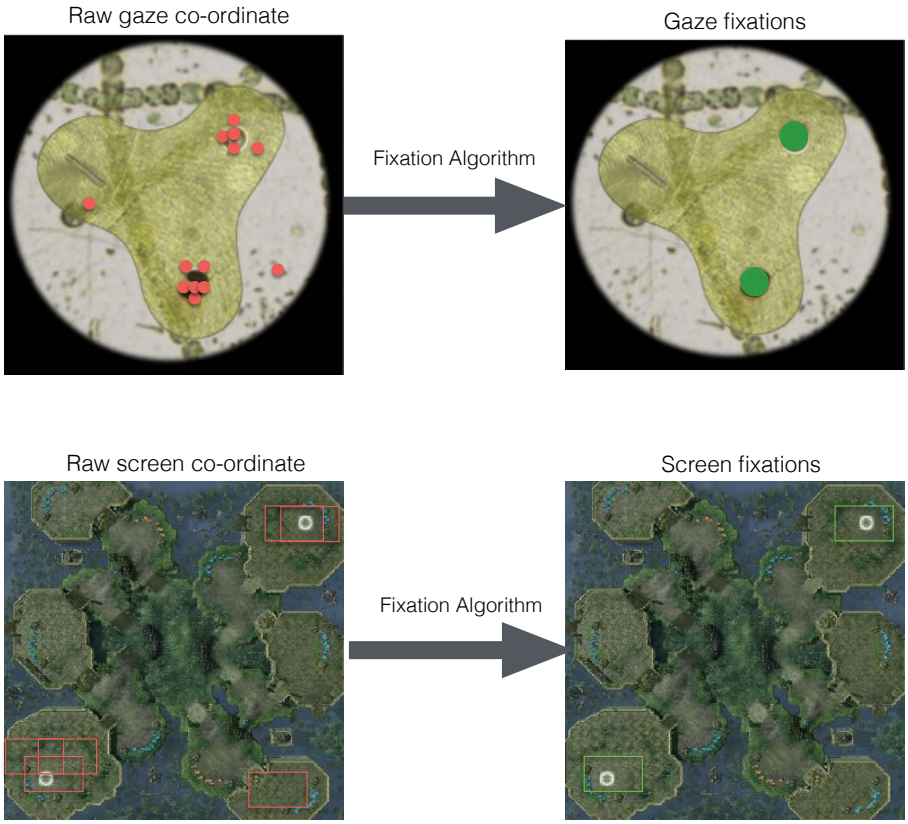


Figure 2.3: Illustration of screen fixations: The same algorithms (Salvucci & Goldberg, 2000) used for aggregating gaze movements into fixations (e.g., top), is employed to aggregate screen movements into screen fixations (bottom).

workers that mine and build military structures. The mechanics for controlling pieces and moving the screen is also the same across all races, so it is possible to compare races quantitatively. Nevertheless, differences between the races makes perfect transfer of learning from one race to another seem rather unlikely. Consequently, race will be an important variable when I consider how to define experience in Chapter Three.

2.2 Skill and the StarCraft 2 Community

The best StarCraft 2 players can safely be called experts. Players compete professionally for prize pools of as much as 1,000,000 dollars (*Esports world championships at Blizzcon 2016*, n.d.), and receive sponsorships from major corporations (*Team Liquid: Personal Sponsorship*, n.d.).

StarCraft 2 professionals often practice full time in 'gaming houses', were players train with other professional players. One of the few peer-reviewed studies on the practice habits of professional e-sport players reports that, contrary to media reports of professional players

training 10-14 hours a day, they practice ab average of 5.28 hours a day, which typically includes about an hour of exercise (Kari & Karhulahti, 2016). One difficulty of this study was that it had a relatively small sample (115 individuals) spanning multiple e-sports, and only 13% of the participants in this study were StarCraft 2 players. A second issue is that it did not report data on top-tier professional players, suggesting that these extreme experts might practice even more. However, these results do fit with data on hundreds of high level amateur players from Thompson et al. (2013), who reported around 3.8 hours a day. Finally, 4 hours of practice per day fits well with classic studies on deliberate practice in violin experts (Ericsson et al., 1993).

Further evidence that StarCraft 2 is a game of skill is that its developer needs to take care matching players against others of similar skill. All those playing competitively online are organized into seven leagues (Bronze, Silver, Gold, Platinum, Diamond, Master, and GrandMaster). Underlying these placements are, at the time of data collection, players' hidden match making ratings (MMR). The primary goal of this system is that players of equal ratings should have an equal chance to win. The algorithm underlying MMR is kept hidden by the game developer, but it has been reported that the algorithm is similar to match-making systems in other video games (*Battle.net Leagues*, n.d.). If so, I can assume that, at the very least, winning matches against better opponents raises or lowers MMR based in part upon the skill of the opponent (Herbrich, Minka, & Graepel, 2007). If StarCraft 2 were not a game of skill, there would be no need to implement such a ranking system.

StarCraft 2 can be played with other formats. Players can compete online in 2v2, 3v3, and 4v4 games, where a team-based MMR allows for matchmaking. However, while StarCraft 2 allows for many game-types, 1v1 performance is the standard of StarCraft 2 expertise. Professional StarCraft 2 is almost exclusively 1v1. It is less clear whether other game formats constitute a domain of expertise in the same way that 1v1 performance does.

2.3 Data Collection

All of the data presented in this work were collected using online surveys and replay file donations. A call for participants was sent out on various websites frequented by StarCraft 2 players, including TeamLiquid.com and Reddit.com/StarCraft. Participants gave informed consent, filled out an online survey, and donated a recent replay file.

The present work relies on two datasets. The first dataset, which I will call the cross-Sectional dataset, was used in previous work (Thompson et al., 2013). It includes 3,307[1] individuals from seven levels of skill. Each participant completed a survey and donated a re-

[1]The sample size differs slightly from the original study because of parsing differences and the recognition that a few individuals were able to submit two replays. These individuals have been dropped.

cent 1v1 replay. The dataset also only contains 29 female participants, so no generalizations will be made to female populations based on these data.

My longitudinal dataset was collected later on. Despite a few minor changes to the survey questions, the big difference between the two studies was that, in the longitudinal study, I was only interested in participants who could donate at least 300 replays. Collecting this many replays is easier than it might sound as StarCraft 2 allows players to save all of their replays by enabling a feature in the game options. Importantly, the survey respondents contained two Female participants and 101 males. One player preferred not to answer, and one player listed their gender as 'Other.' Consequently both datasets lack information on non-male players, so I will restrict any generalizations to males.

2.4 Flow of Data Processing

When a StarCraft 2 game is played, users are allowed to save their game in a 'replay file.' The main purpose of these files is to allow users to watch their previous games and improve. However, replay information is not stored as a collection of pixels like a video might be. Instead, replay files contain records of every action performed by a player within a game. This allows entire games to be reconstructed by using the StarCraft 2 engine and a small log file.

I used the software SC2Gears (Belicza, 2014) to unparse replay files into three text files. One file a series of actions within the game, a second included metadata about the game (including information such as the date of the game), and a third contained data about player chat. These data were then uploaded to MySQL tables. A series of Python and Matlab scripts aggregated these data into higher levels of analysis. This includes aggregating the screen movements of individuals into screen fixations, and calculating key variables.

Finally, a series of Matlab scripts checked the database for consistency and rooted out games which needed to be excluded. Out of the original 164,001 games that reached the game level, I excluded 44,589 meaningless games whose data I cannot interpret. These are games which failed a test of data integrity (0 games), could not be paired with a survey respondent (16,801 games), were associated with multiple survey respondents (9,619 games), appeared to possess impossible game versions (29 games), appeared to be duplicates based on a file size identifier (1 game), or appeared to be played by the same players, on the same date, and time (8 games). An additional 35,835 were dropped because the survey player's character name could not be identified in the game, and 5,240 were dropped for having fewer than 100 rows of raw data (typical games contain many thousands of rows).

Secondly, an additional 35,982 games were judged as meaningful but were not in the format that is indicative of StarCraft 2 expertise. Across the entire 164,001 games of the dataset (which includes meaningless games), 53,261 had the wrong format (games which were not played on 'faster' speed, were not 1v1 games, or were not matched against an

opponent online using Blizzards match making system). Furthermore, 8,834 were dropped for involving at least one computer player, 32,570 were dropped for having the wrong number of clients associated with the game[2], and 9,874 were dropped because they appeared to contain a third-party observer. Finally, one game, which passed all the required tests, was missing from the final dataset due to human error, and one player was dropped for having only five games contribute an observation to the dependent variable.[3]

I do not attempt to explain performance within these games (i.e., they do not constitute an observation on the DV), but they are reflected in a player's history (e.g., in the number of 2 versus 2 games played). This resulted in a dataset of 83,429 games (about 20,850 hours of performance) across 107 individuals. An additional game was dropped because the survey player did not register any screen fixations.[4]

2.4.1 Lag and Resolution

StarCraft 2 is a multiplayer game, so the current game state must be updated across all players. Over the course of a second of play time, the StarCraft 2 engine ensures synchrony a number of times. At these transition periods the actions of individuals receive a timestamp based on when the game state is synchronized. While the details of this system are not given by the game manufacturer and likely change over time, it is clear that they lead to somewhat discontinuous latency data. For example, in the skillcraft database of 996,163 fixations (Thompson, McColeman, et al., 2017) with at least one action, the smallest latencies are noticeably bunched up. There are 54,534 first actions with a latency of zero (which are either extremely fast or brought about by a lagged network connection and delayed synchronization), 21,853 games with a latency of 90 milliseconds, 98,371 first actions of 135 milliseconds, and 47,124 have a latency of 226 milliseconds. Given that these discrete jumps can come close to 100 milliseconds, I need to be careful about interpretation of an individual speed measurement. It also means that there will be disadvantages of both the mean and median when aggregating across a single games performance data. The means are potentially problematic because latencies will exhibit a floor effect, leading to values that might be heavily impacted by outliers. Medians are problematic because they will greatly reduce the number of unique values for any given measure (e.g., 10% of the games may

[2]In the replay file, a game client could be a player or an observer who watches the game without participating. A 1v1 game which is set up using Blizzards match making system will have exactly two clients.

[3]I take the one missing game to be completely missing at random because every game appears to have had an equal chance of being missed in this way. I do not bother with imputation because it is only one game.

[4]It is important to note that it is possible that I have not sampled all the games from each player, especially older games from early in the player's career. I need to acknowledge that individuals will likely start with different intercepts, perhaps because of prior experience with other real-time strategy games or perhaps even prior experience with StarCraft 2.

end up taking on the same value). Here I generally give preference to medians despite its disadvantages because there are still many unique values of FAL.

Chapter 3

What is Experience?

In laboratory studies, a workable operationalization of experience is usually freely available and obvious. For example, if one is running a learning study with exactly one exposure to a stimulus per trial, then it usually makes sense to measure experience by the number of trials.

In cases where the dependent variable is sampled throughout the entire experiment (such as measures derived from eye tracking), there may be a choice as to whether to measure experience in terms of trial number or time. The choice matters somewhat, for example, in tasks where trial duration decreases with experience. One definition may be more convenient than another, or yield easier interpretation. But as Chapter Two makes clear, StarCraft 2 contains a number of game modes and races which may or may not transfer. One might argue that similar difficulties arise for any complex skill.

It is worth the effort to contemplate the distinction between transfer and learning, given that both imply a change in behavioural capacities based on experience. One might be able to keep questions of transfer and learning separate if there were clear boundaries for what constituted domain performance. In this case, learning would refer to skills acquired from intra-domain experience, while transfer would refer to skills acquired from extra-domain experience.[1] For example, the present work can sidestep the question of how basketball skill transfers to StarCraft 2. Basketball is so far outside of the domain of StarCraft 2 that one can sensibly study learning in StarCraft 2 while acknowledging that an understanding of transfer will eventually be needed in any complete theory of StarCraft 2 performance. The problem is that in complex environments, the boundaries of what constitutes StarCraft 2 are somewhat fuzzy. Games with four StarCraft 2 players are almost never seen in professional StarCraft 2 (and it is not possible to be full-time professional 2v2 StarCraft 2 player), but the close relationships between the game modes are obvious to anyone experienced in the

[1]This idea is owed to helpful conversations with Kathleen Slaney.

game.[2] It is also unclear whether I can say that a professional Terran player who, for the sake of a charity event, plays as the Zerg race, is still playing the same game. Given the remarkable consistency in which professionals select their races, there is an argument to be made that expertise should be restricted to a chosen game race. It is not obvious, for example, that a 'Terran' professional who switches to 'Zerg' overnight *is still an expert in StarCraft 2*.

Since the boundaries of what constitutes competitive StarCraft 2 are so fuzzy, I do not make any effort to distinguish between the *intra-domain learning effects* and *extra-domain transfer effects*[3]. What is important is that the resemblance between the different kinds of StarCraft 2 are so strong that I must address the question of whether to include or exclude these sorts of data in an analysis of performance. Unlike the question of whether basketball transfers to StarCraft 2 performance, which is understandably put aside for future work, decisions about what to include in the operationalization of StarCraft 2 experience could impact results substantially.

Operationalizations of experience can be too broad or narrow, and either mistake could undermine an analysis of performance. If they are overly broad, learning curves can form discontinuities as participants pile up heaps of experience that are not relevant to performance. If the definition is too narrow, then I might expect inexplicable leaps in performance as participants draw on experiences which researchers have overlooked. Both problems could obfuscate the effect of experience on performance. Furthermore, since the shapes of learning curves are of theoretical relevance in themselves (Bryan & Harter, 1897, 1899; Gray & Lindstedt, 2016; Keller, 1958), the presence of potentially artifactual discontinuities is a serious methodological problem.

In the next sections I will address an approach for dealing with this methodological problem. For now, it is important to be clear that evaluating claims about transfer will need to be, at least in part, an exploratory analysis. This is because facts about which skills will transfer in a new domain are poorly understood and often counter-intuitive. This makes it difficult to formulate predictions about when transfer will be observed.

[2]It is worth emphasizing that the game rules are consistent across these game modes. The only difference is the number of players on each team. This difference, however, can have pronounced consequences for what constitutes an adaptive strategy. For example, the combined forces of two players will usually be able to overwhelm a single opponent early in the game, making it important for team mates who are positioned far apart to be ready to support each other on short notice. Websites offering advice to players who want to play 2v2 competitively often acknowledge that this sort of 'early pressure' is more effective in 2v2 games (Andress, n.d.). At least one such site (Andress, n.d.) points out that the success of early pressure may change the nature of the StarCraft 2 information access problem. It is often useful to scout the opponent's base to ascertain their strategies. In 2v2 games, they argue that scouting is more important because it allows a team to station both of their armies in locations that will receive early pressure.

[3]Recall that Wittgenstein (1953) used games as an example of classes for which membership needs to be thought of in terms of resemblance rather than necessary and sufficient conditions. If he was right, providing necessary and sufficient conditions for what constitutes StarCraft 2 may be similarly futile, even though all the game modes of StarCraft 2 are played within the official 'StarCraft 2' game software.

3.1 Variables

Because the vast majority of professional StarCraft 2 games are 1v1, it makes sense to treat 1v1 performance as the primary explanandum for a theory of skill learning. My data will contain one observation for each meaningful 1v1 game a player has.

- First Action Latency Medians ($medianFAL$): A First Action Latency is the time between the onset of a screen fixation and an action (also see Chapter 2.1.1). I take the median First Action Latency for every 1v1 competitive StarCraft 2 game. It is useful to note that, because of issues with lag and resolution discussed in Chapter Two, median FAL will have fewer distinct values than I would like (see Table 3.1).
- $1v1_{xp}$: Since I am trying to explain 1v1 performance, it makes sense to begin from the assumption that the best measure of skill is the number of competitive 1v1 games played.
- *Race*: a nuisance factor indicating whether the player chooses to play the Terran, Zerg, or Protoss race. Some races tend to be faster than others (Thompson, McColeman, et al., 2017).
- Pr_{xp}, Tr_{xp}, Ze_{xp}: indicators of the number of 1v1 competitive games (games where teams are matched online using the developer's skill-based match making system) played as a particular race. These variables allow for more narrow definitions of experience as being specific to a particular race.
- $2v2_{xp}$: indicators of the number of competitive games (games where teams are matched online) played with exactly two opponents on each team. This is the second most common format for a game, with 1v1 games as the most common.
- $N-2v2_{xp}$: the number of competitive games (games where teams are matched online) that are neither of the 1v1 or 2v2 format. This includes 3v3 games, 4v4 games, custom games (which include other formats), and Free-for-all games. From Table 3.2 it is clear that the most common formats in this category are 3v3 and 4v4 games. The lone custom game was listed by my system as having five players, which may have occurred because a player left the game before having actions logged into the replay file.

3.2 The Distribution of Skills in the Longitudinal Sample

It is sensible to get a sense of the distribution of skill levels in my longitudinal data before addressing my research questions. It would be useful to know, for example, if my participants were predominantly skilled players.

median FAL (raw timestamp)	Frequency
	V1
0	2
10	1
12	207
16	133
18	7
20	1156
22	54
24	4018
26	49
28	1734
30	59
32	6461
34	276
36	9085
38	109
40	2476
42	120
44	22586
46	653
48	8011
other	26232

Table 3.1: Unique median FAL values.

Unique game formats	
count	format
102,549 (83,429)	1v1
9,031 (8,452)	2v2
3,882 (3,706)	3v3
2,955 (2,115)	4v4
201 (1)	Custom
791 (632)	FFA
3 (0)	Unknown

Table 3.2: Counts of meaningful StarCraft 2 games by format. Note that only a subset of 1v1 games satisfy my definition of true competitive starcraft and thus contribute an observation to the dependent variable (see 2.4 for details). Also note that games with more than two players only will count as experience if the game was confirmed to be a competitive game (see 2.4 for details). The number of games that count towards an individuals potential experience is listed in parentheses.

Although I have access to some MMR data, it is not ideal for this purpose as this data only covers a small number of participants, and it only estimates their skill after the release of 'StarCraft 2: Heart of the Swarm'. Instead, I compare participant mean FALs[4] against the mean FALs typically observed in the cross-sectional dataset. Note that I use *mean* FALs rather than median FALs for this purpose as I only use performance measures to get a rough indication of player skill. The mean FALs are more appropriate for this purpose given their appearance in previous work (Thompson et al., 2013), while median FALs appear more appropriate as a non-biased indicator of the central-tendency of a player's performance latencies.

I employed a sample of 82,534 out of the total 83,429 possible games for this sample because of the unavailability of both median and mean FALs.[5] Mean FALs tended to be about 141.5 milliseconds longer than median FAL due to the presence of outliers in the distribution of First Action Latencies (see Figure 3.1).[6]

Figure 3.2 shows how the mean FALs of players in the longitudinal study stack up against the mean FALs of players from the cross-sectional dataset. It initially appears that bronze and silver players may be somewhat underrepresented in the dataset here. However, a look at Figure 3.3, which shows the first fifty games from each player, suggests a more balanced sample of skills. The apparent underrepresentation of novices in Figure 3.2 may therefore be partially due to early learning effects.

3.3 Analytic Strategy

I want to explain 1v1 performance, so my initial approach will be to assume that the best operationalization of experience is direct experience with the target domain. So my presumptive definition of experience is the 'number of 1v1 games played.' Since the presumptive definition could be misleading, I then look for evidence in favour of more restrictive definitions (such as Pr_{xp}, Tr_{xp}, and Ze_{xp}). Finally I will examine whether the definition also needs broadening (perhaps by including experience from other game formats such as $2v2_{xp}$ and $N-2v2_{xp}$). The analytic strategy will be as follows.

1. Begin with a prior operationalization of experience which directly matches the target domain ($1v1_{xp}$).

[4]Recall that Thompson et al. (2013) identified mean FAL as a robust predictor of skill in StarCraft 2.

[5]Missing games were missing due to human error in the parallel processing of mean FALs, so I treat them as completely at random and drop them from the dataset for this analysis.

[6]It is worth noting that Figure 3.1 compares mean *mean FALs* from a cross-sectional dataset to median *mean FALs* from the longitudinal dataset (These correspond to the bars at the centres of each box).

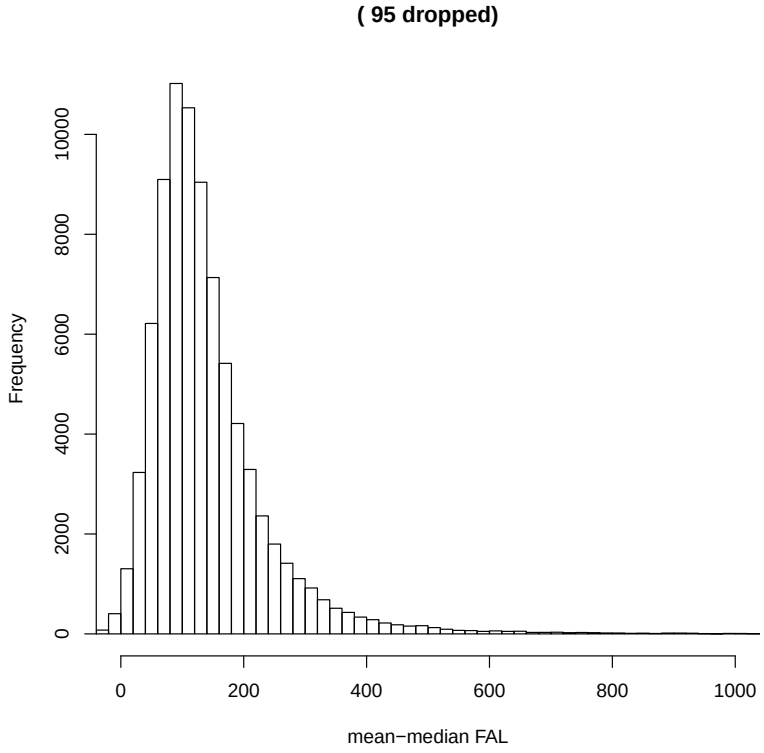


Figure 3.1: Histogram of the differences between mean and median FAL. It is important to note that 95 extreme differences are dropped for the purposes of visualization, and that data are not independent (there are multiple observations per player). These plots are based on a subsample of the dataset with only 82,534 games out of the total 83,429.

2. Construct a presumptive model relying on this definition of experience. Modify exclusion criteria as necessary.

3. Test the presumptive model against more complex models which include multiple measures of experience, including experience in particular roles (i.e., models with Pr_{xp}, Tr_{xp}, and Ze_{xp}). This will be done with a series of likelihood ratio tests on nested models.

4. Revise the presumptive model in light of evidence

5. Test the presumptive model against more complex models that include extra-domain experience, such as $2v2_{xp}$ and $N - 2v2_{xp}$.

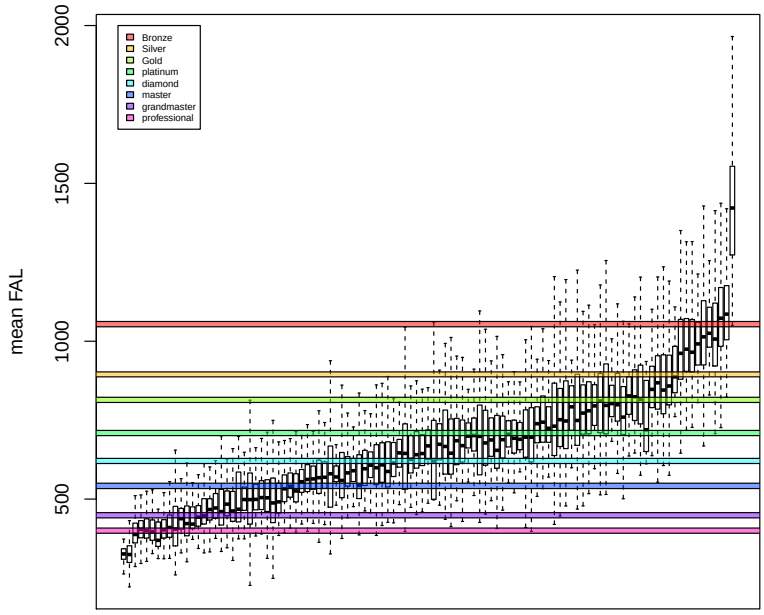


Figure 3.2: Boxplots of mean FAL for each player in the dataset, with outliers removed (data points greater than 1.5 times the interquantile range from the third quantile or less than 1.5 times the interquantile range from the first quantile). See Figure A.1 to see this figure with outliers included. Transparent bars are centered around the median *mean FALs* from Thompson et al. (2013). The width of the bar reflect 2 times the standard error of the median for the league in question (acquired from the standard deviation of 1,000 bootstrapped subsamples, Mangiafico, 2016; Ripley, 2017). The reader should assume that the eight distributions of mean FAL in Thompson et al. (2013) overlap considerably. Finally, these plots are based on a subsample of the dataset with only 82,534 games out of the total 83,429.

3.4 Selecting a Presumptive Model

3.4.1 Raw Data

Figure **??** highlights the importance of considering race differences, outliers, and small samples. The starkest differences in slope here are seen within player and across race. These radical differences are presumably brought on because certain players avoid certain races, leading to small sample sizes and highly variable regression lines. Also note that there are some extreme outliers in median First Action Latency which might impact regression lines. Indeed some values are more than 10 standard deviations from the mean. Finally, it should be noted that I simply lack data for some players (e.g., player 119), making interpretation

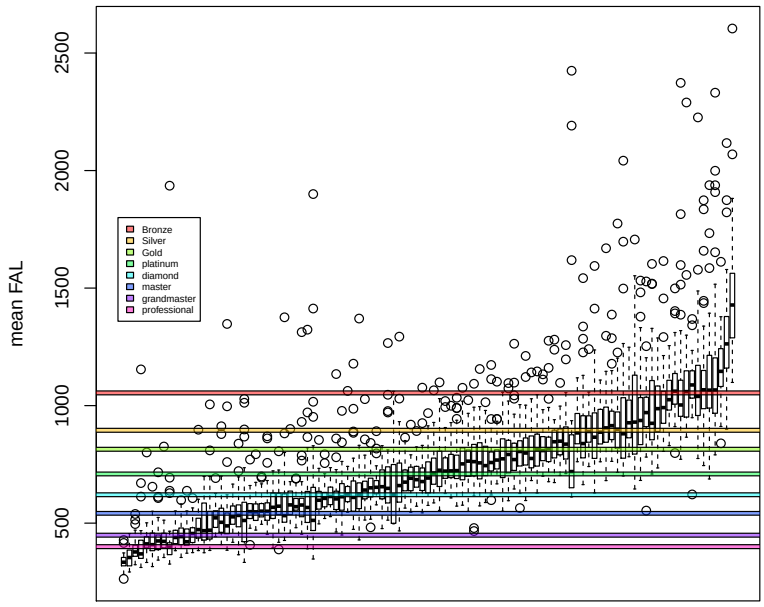


Figure 3.3: Boxplots of mean FAL of the first fifty games for each player in the dataset. Transparent bars are centered around the median *mean FALs* from Thompson et al. (2013). The width of the bar reflect 2 times the standard error of the median for the league in question (acquired from the standard deviation of 1,000 bootstrapped subsamples, Mangiafico, 2016; Ripley, 2017). The reader should assume that the eight distributions of mean FAL in Thompson et al. (2013) overlap considerably. Finally, these plots are based on a subsample of the dataset with only 82,534 games out of the total 83,429.

difficult. Before specifying the presumptive model, therefore, I address the following three tasks:

1. investigate and remove of problematic outliers
2. deal with methodological issues associated with race data
3. consider removing players from the analysis who only have a small number of games, and consider removing games from players who play some races very infrequently

3.4.2 Race

The sample contains a respectable number of games from each race, with about 39 percent of games played by Protoss, 28 percent Zerg, and 33 percent Terran games (see Figure 3.6).

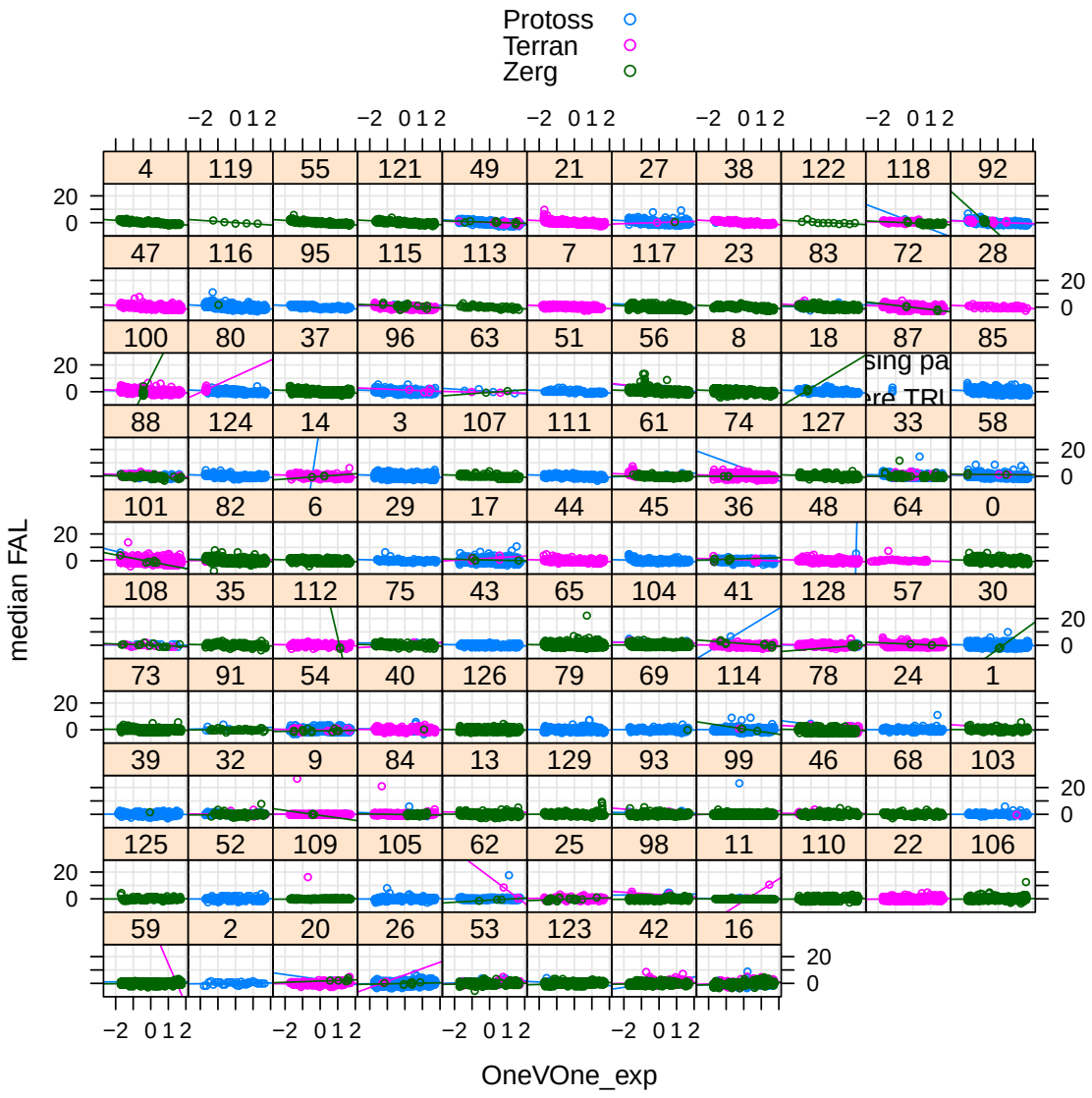


Figure 3.4: Initial Lattice Plot for data. Each players median FALs are normalized individually. Researchers should note that this transformation makes it easier to observe relationships between experience and median FAL, but impossible to gauge differences in the intercepts among players.

This distribution is only slightly off from what is typically observed in websites authored by StarCraft 2 fans which keep track of the number of players by race. Rankedftw.com, for example, typically reports around 30 percent of games being played by Protoss players (*Ranked For Teh Win*, n.d.).[7]

[7]Thanks goes to Tim Racine for pointing out that, despite a number of arguments in the community about which race is superior overall, no race appears to be preferred by an overwhelming majority of participants.

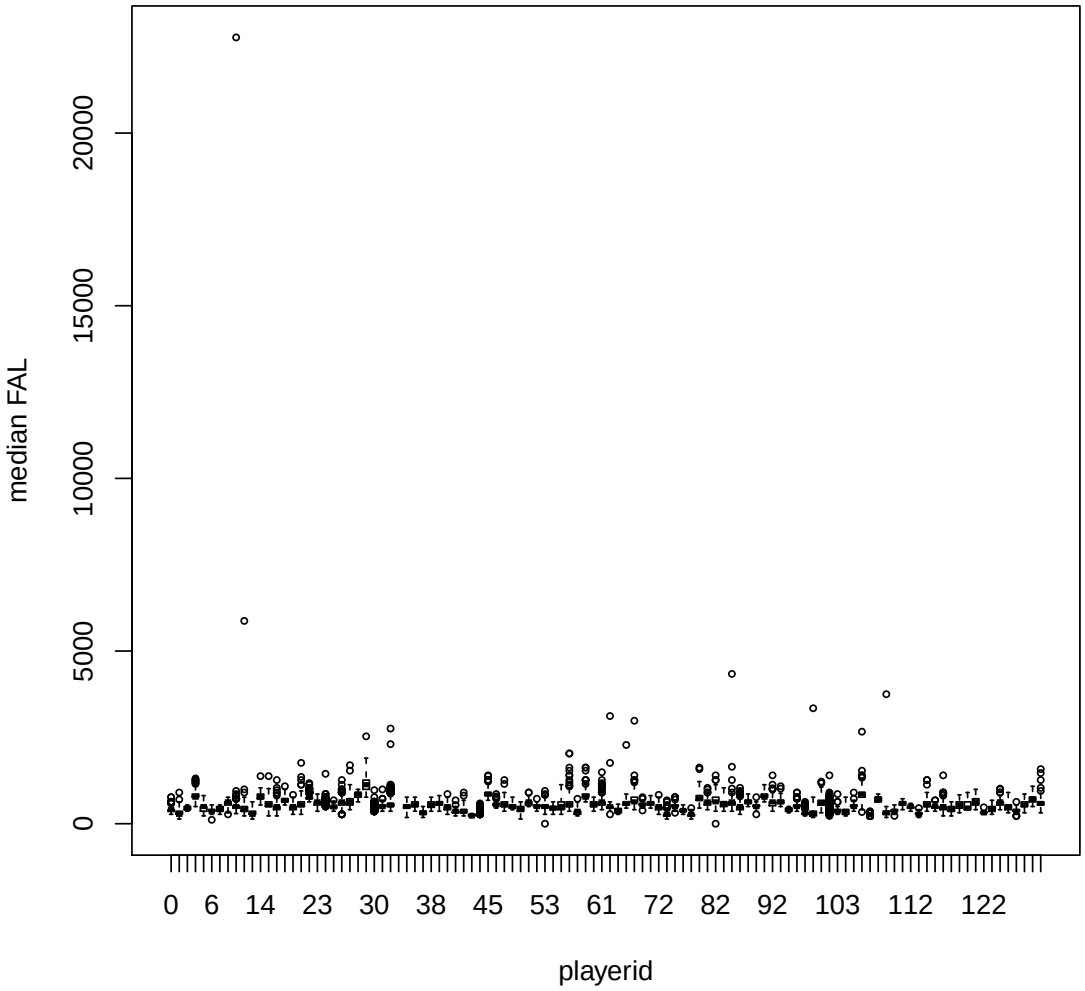


Figure 3.5: Boxplot for median FAL by player.

A look at Figure 3.7 confirms that many of the frequently playing participants play Protoss almost exclusively (with four participants playing over 3,000 games each). It is also clear that players who play as Terran are unrepresented at the highest levels of experience (3,000 games and up), so I will be especially cautious in making interpretations about players of this sort. Another concern is that a large proportion of the Terran games come from a single, exceptional participant who appears to have played over 7,000 games of StarCraft 2. I will need to be cautious when interpreting models to ensure that such individuals are not given undue influence.

It is clear that vast majority of players in the sample choose a dominant-race and spend the majority of their time playing the same race. I refer to these participants as sticking to

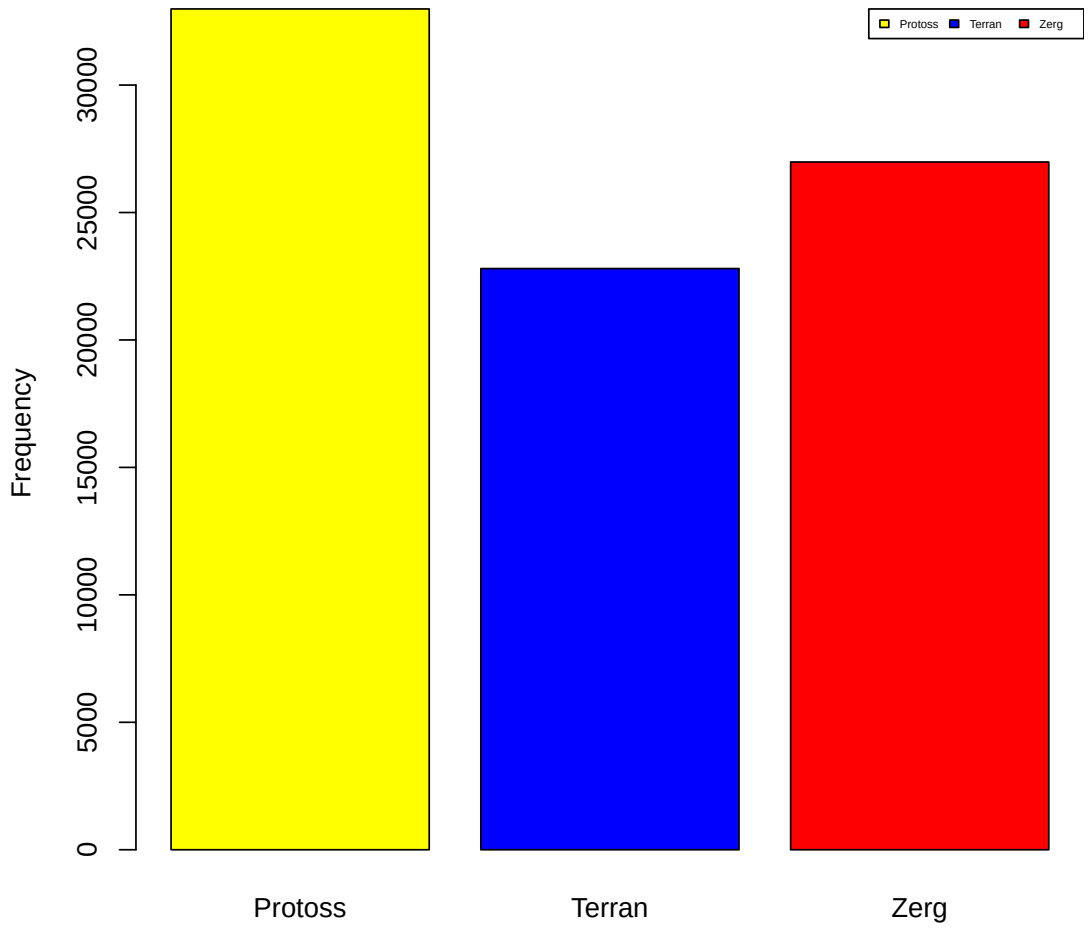


Figure 3.6: The number of games for each Race. Data are not independent (each player contributes multiple observations).

a 'dominant-race', which I simply define as the most common race played by a participant. I describe a player as playing 'off-race' when they play StarCraft 2 without using their dominant-race.

It is important to recognize that race could impact performance in two different ways, both of which require investigation. First, some races might allow for faster game-play than others. I address this question in the construction of a presumptive model, as it only requires thinking about race as a simple between-subject that influences a player's overall speed in a fixed manner. The second possibility is that learning is race-specific (e.g., dominant-race experience might only contribute to performance in dominant-race games). The latter

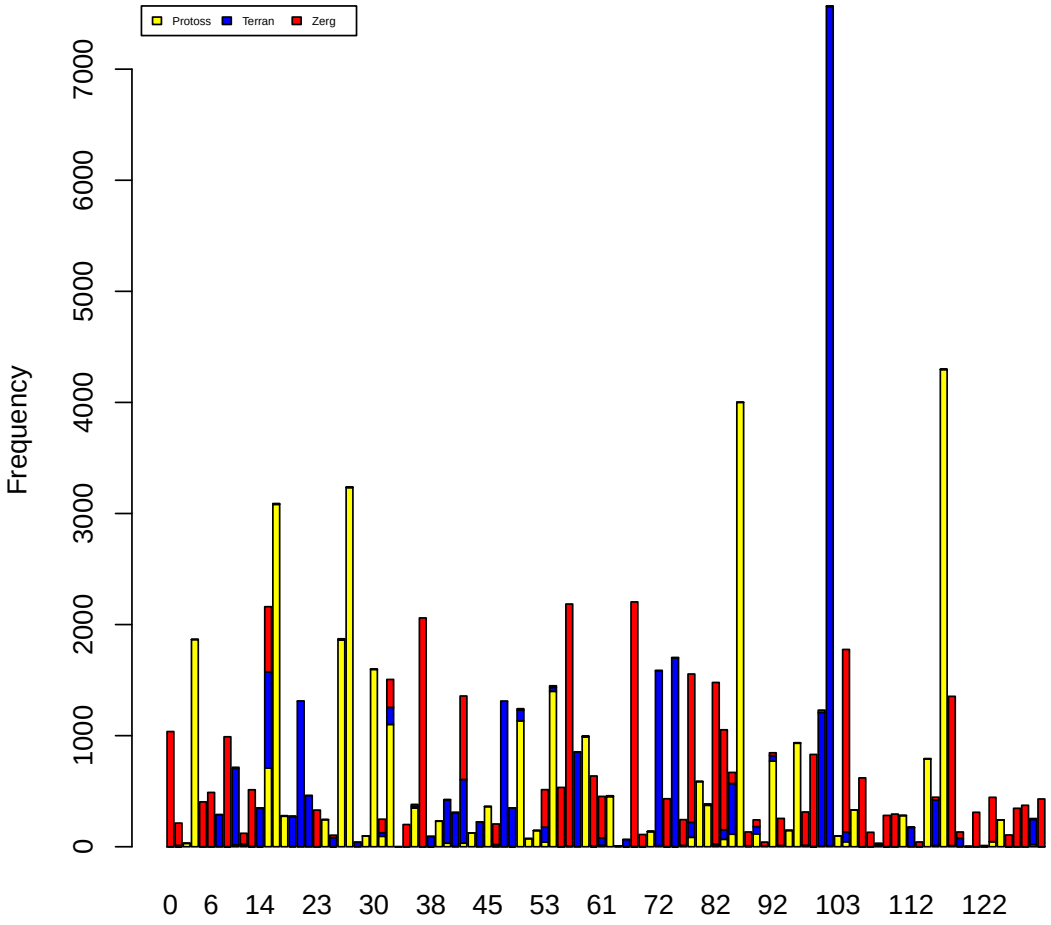


Figure 3.7: Distribution of Race data by player.

question is one of transfer, and it therefore is a challenge to the presumptive definition of experience which views all standard 1v1 StarCraft games as being, on average, of equal relevance to learning. I consequently address it only after having established a presumptive model.

3.4.3 Dropping Participants with Very Small Quantities of Data

I will ultimately want to compare the presumptive model to ones which have effects associated with race-specific experience, and even interactions between experience with a race, and a variable indicating the current race being played (e.g., a history of Zerg experience

may only benefit a player who happens to be playing Zerg in a given game). Consequently, players with very few off-race games will be problematic in constructing such models.

I dropped off-race data for a player if they have fewer than 30 games of that race. As a result of this criterion, I dropped an additional 656 games from 61 different players. Four players were dropped entirely, for failing to have more than 30 games from any race.[8]

3.4.4 Dealing with Outliers

Figure 3.4 reveals a large number of outliers that are likely to have undue influence on models and visualizations. Some of these outliers are so extreme that I begin to doubt they are representative of actual gameplay. Players might be playing under additional distractions (e.g., watching television), carrying on conversations, or frequently leaving the keyboard for a walk to the fridge. Figure 3.5 gives a clearer depiction of some of the most extreme outliers. I exclude 53 games with a median FAL that is more than 3 interquartile ranges from that player's interquartile range. I suspect that players in these cases are likely playing under such unusual circumstances (e.g., extreme distraction) that I do not attempt to account for them here.

3.4.5 Summary of Data Exclusion

After dropping data 103 players and 82,768 games remain. Figure 3.8 shows the resulting dataset (scaled after exclusion). To make visualization possible, 73 extreme data points more than five standard deviations from the mean have been ignored (but not dropped from eventual analysis[9]). This finally allows one to observe the impact of experience on response latencies.[10] If I disregard race, the first four rows of 44 players certainly appear to gradually speed up over experience, while only a handful (perhaps the bottom 7) seem to slow down with experience.

3.4.6 Specification of the Presumptive Model

So far I have only stated that the presumptive model should accept the tacit assumption that $1v1_{xp}$ is a workable operational definition of experience. The structure of the presumptive model still requires a full specification.

Figure 3.5 makes it clear that players do vary substantially, and since I think of participant ID reflects a random sample from a population of possible participants, rather than

[8]Note that if a game is dropped, we will not attempt to explain performance within that game (i.e., it does not contribute an observation to the dv). However, our measures of experience will still log the games existence (i.e., future behaviour may be explainable based on these experiences).

[9]Plots of the entire dataset will be presented elsewhere.

[10]But the reader should take care given that these are standardized data, obfuscating differences in intercepts (which may be related to slopes in learning data).

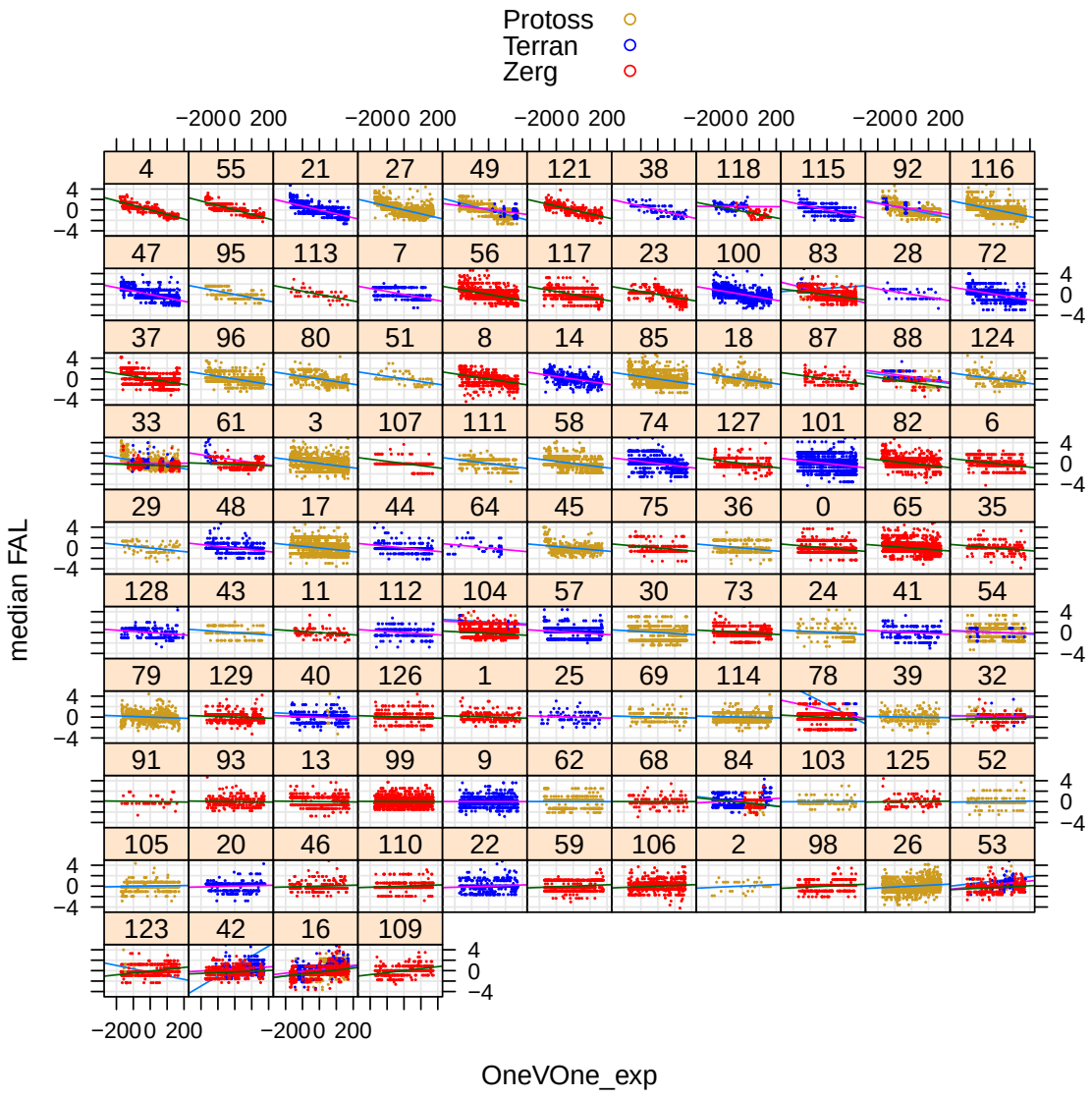


Figure 3.8: Median FALs by $1v1_{xp}$, representing data from Figure 3.4 after exclusion of 680 games from 59 players. In order to the for visualization of the bulk of the data, the y axis still excludes 71 extreme data points that remain in the analysis (plots containing the entire dataset will be shown after the construction of the best model). Players are ordered by the slope coefficients resulting from simple linear regressions of median FAL on $1v1_{xp}$.

a exhaustive list of possible factor levels, it makes good theoretical and practical sense to construe participant ID as a random effect rather than a fixed effect (D. M. Bates, 2010).

The simplest structure I consider is:

$$Model_1 : FAL \sim 1v1_{xp} + (1|Participant_{ID})$$

Which contains a fixed effect for direct experience and includes a random intercept allowing participants to vary in their starting median First Action Latency.

To complete the presumptive model, I need to make decisions about:

1. Does a player's race have a direct impact on their speed within a game (i.e., is there something about the domain allowing some races to play faster than others)?
2. Should I include a random slope for $1v1_{xp}$ (i.e., should I think of different participants have having different learning rates)?

I make the first decision by comparing Model 1 to,

$$Model_2 : FAL \sim 1v1_{xp} + Race + (1|Participant_{ID})$$

Since Model 1 is nested within model 2, a likelihood ratio test is appropriate. I will follow a similar strategy throughout this chapter.

Decision 2 requires comparing $Model_1$ to $Model_3$:

$$Model_3 : FAL \sim 1v1_{xp} + Race + (1 + 1v1_{xp}|Participant_{ID})$$

Errors approximated a normal distribution, and residual plots did not reveal heteroscedasticity or non-linearity. For example, residual plots for all models of the models were very similar to Figure 3.9. I have few concerns about heteroscedasticity here, and take the thicker portions of the plot to reflect predicted values which are associated with more observations.[11] There is a small gap in the residual plot where the model avoids making predictions (i.e., the 950 millisecond region), though it is important to recognize that 0.6% of the data comes from games with a median FAL of more than 950 milliseconds. It is worth having a healthy distrust of model predictions at these extremes. Boxplots of residuals by $1v1_{xp}$ and *race* reveal no heteroscedasticity.

The likelihood ratio test[12] suggested that the superiority of $Model_2$ over $Model_1$

(Model_1_{df}=4, $Model\_2_{df}$ =6, $\chi^2(2)$=1166.26, $p < 0.0001$).

This confirms that race actually impacts performance measures, a finding that is also supported by prior research (Thompson, McColeman, et al., 2017), and by a qualitative look at how the game mechanics vary by race.

$Model_3$ which adds a random slope to the model, yielded a better model still.

(Model_1_{df}=6, $Model\_2_{df}$ =8, $\chi^2(2)$=14521.55, $p < 0.0001$)

[11] There was also no indication of heteroscedasticity in plots of residuals against participant ID and other predictors.

[12] for the purposes of likelihood ratio tests models are refit with maximum likelihood estimation and for parameter estimates restricted maximum likelihood is used. Full output from likelihood ratio tests can be found in Appendix ii.1. This follows D. M. Bates (2010).

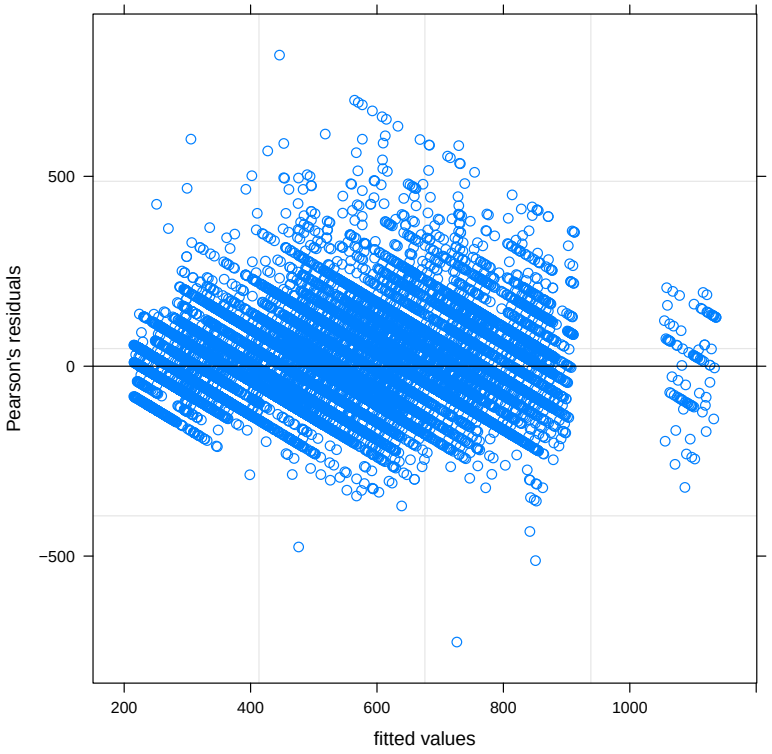


Figure 3.9: Residual plot for Model 3 using the method described in the R package 'Lme4' (Douglas et al., 2017).

It seems increasingly likely that some individuals learn more than others. While the best model does not appear to violate any assumptions of the likelihood ratio test, I explore the workings of this model before turning to my primary goal of comparing this presumptive model to those introducing more nuanced definitions of experience.

Figures 3.4.6 shows how model performance falls when altering parameter values (D. M. Bates, 2010).[13] For a complete summary of the presumptive model, see Appendix ii.1. The dependent variable, ζ, is the signed root of the likelihood ratio statistic (D. M. Bates, 2010), allowing for comparisons of the optimal model with non-optimal parameter settings. The out-most vertical lines in 3.4.6 can be interpreted as 99% percent confidence intervals on parameters (D. M. Bates, 2010). According to this model, StarCraft 2 Protoss players are expected to begin their career with First Action Latencies of about 570 milliseconds (95% CI [540,610]). Terran players tend to be roughly 22 milliseconds slower while Zerg players tend to be about 44 milliseconds faster. When a player continues playing with the same

[13]Following D. M. Bates (2010), I also looked at profile pair plots to rule out the possibility that the model performance changes across one parameter change sharply at slightly different values of a second parameter. I do not include them as Figure 3.4.6 appears representative.

race, the presumptive model predicts an improvement of about 20 milliseconds every one hundred games.

One might also note the moderate negative correlation between the random slopes and intercepts of -0.38. Players who start off slow tend to have more negative slopes. This is not entirely surprising as there are likely to be human limits on how fast one can play StarCraft 2, even with extreme expertise. Strong players have less of an opportunity to shave more time off of their already impressive speed.

The first finding of note is that there is great variation in the learning rates of participants (see Figure 3.4.6). Consideration of the standard deviation of random intercept (σ_1) reveals that it should not be uncommon to find individuals starting off 160 milliseconds faster or slower than others. Perhaps more startlingly, the expectation of a 20 millisecond improvement every 100 games will not be manifested for many players, as revealed by the large standard deviation of random slopes, σ_3. While the model does predict that players will generally improve over time, it also predicts that it will not be uncommon to find players who don't improve at all. This is confirmed by Figure 3.8.

This analysis also confirms that a player's chosen race is relevant to their performance within a specific game. What remains to be seen is whether one can understand an individual's *learning* in StarCraft 2 without paying attention to their history of race selection.

3.5 Is the Presumptive Model Overly Inclusive?

An important question about the presumptive model is whether it is wrong to aggregate experience with different races into a uni-dimensional definition of experience. To evaluate whether this aggregation results in the loss of information that can be used to predict and explain learning, I consider three additional models which contain race-specific experience measures.

1. $Model_3^{Pr} = FAL \sim 1v1_{xp} + Race + Pr_{xp} + (1 + 1v1_{xp}|Participant_{ID})$
2. $Model_3^{Pr,Tr} = FAL \sim 1v1_{xp} + Race + Pr_{xp} + Tr_{xp} + (1 + 1v1_{xp}|Participant_{ID})$
3. $Model_3^{Pr,Tr,Interaction} = FAL \sim 1v1_{xp} + Race + Pr_{xp} + Tr_{xp} + Race : Pr_{xp} + Race : Tr_{xp} + Race : 1v1_{xp} + (1 + 1v1_{xp}|Participant_{ID})$

The advantage of this procedure is that the presumptive model is nested within all of the race specific models. This allows the use of strait-forward likelihood ratio tests to evaluate

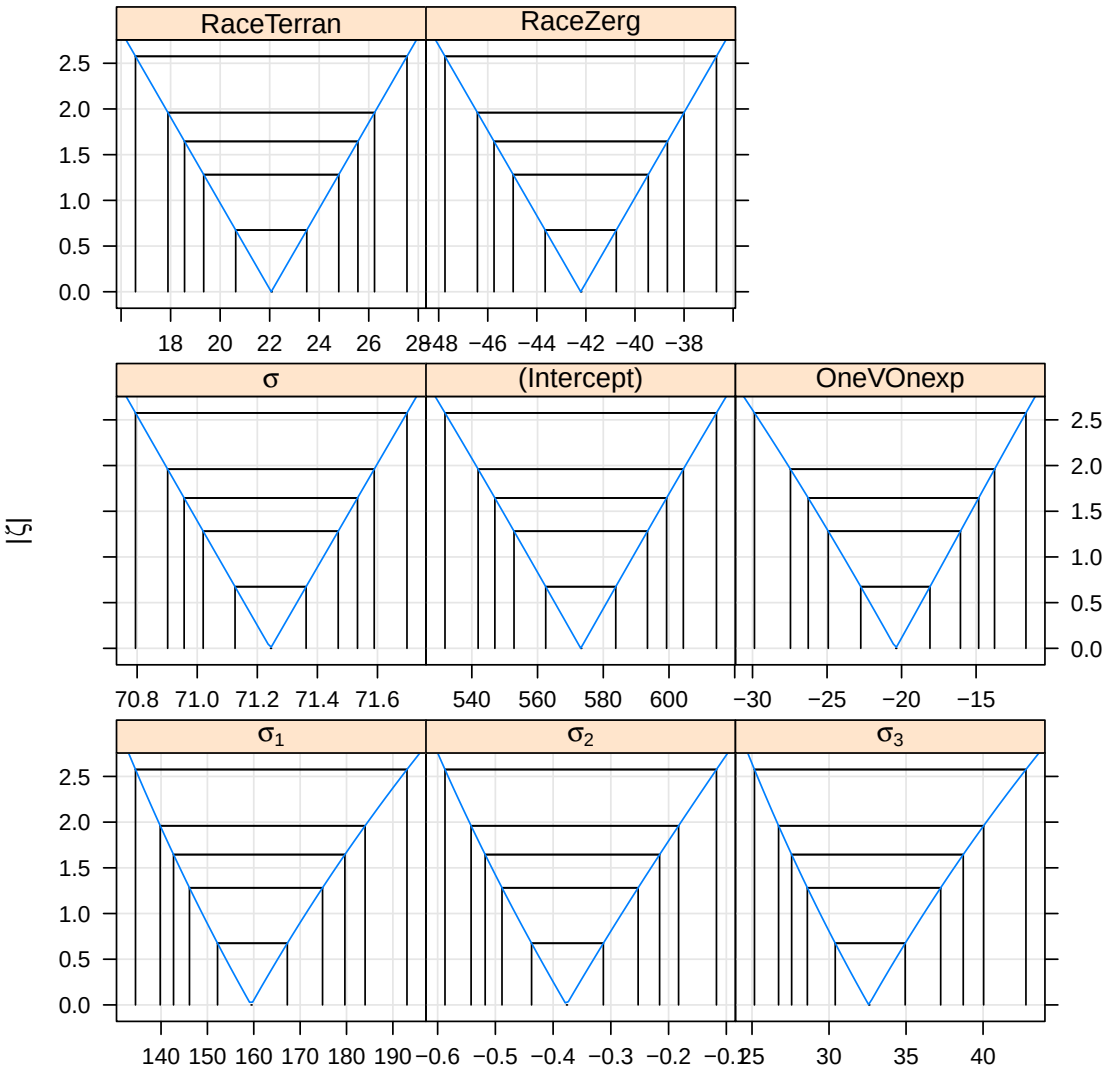


Figure 3.10:
Profile plot of the presumptive model. Boundaries reflect (from outside to inside) 99%, 95%,90%, 80%, and 50% confidence intervals respectively. σ refers to the standard deviation of residuals, σ_1 to the standard deviation of the random intercept, σ_3 to the standard deviation of the random slope, and σ_2 , to the correlation between the random slope and intercept.

whether adding race specific experience variables, specifically Pr_{xp} and Tr_{xp}[14], improves overall model performance.[15]

The disadvantage of this procedure is that the experience measures, $1v1_{xp}$, Pr_{xp}, and Tr_{xp} will be correlated, making parameter estimates of the fixed effects misleading, and the models more difficult to interpret. I sidestep these disadvantages here by using these race-specific experience models only to evaluate whether the intuitive uni-dimensional definition of experience glosses over useful predictive information in race-specific experience. Depending on the success of these race-specific models, I may decide that a uni-dimensional operationalization of experience is problematic. If I do, then it may make sense, given that only a tiny proportion of the dataset contains off-race games, to restrict my analysis goals to the prediction and explanation of *dominant-race* performance.

I found that $Model_3^{Pr}$ was superior to the presumptive model (Model_1_{df}=8, $Model\_2_{df}$ =9, $\chi^2(1)$=23.37, $p < 0.0001$), and $Model_3^{Pr,Tr}$ was superior to $Model_3^{Pr}$ (Model_1_{df}=9, $Model\_2_{df}$ =10, $\chi^2(1)$=14.41, $p = 0.0001$). Race specific experience therefore seemed to generally improve model performance, at the cost of high correlations between predictors. Perhaps more troublesome was $Model_3^{Pr,Tr,Interaction}$ which beat out $Model_3^{Pr,Tr}$ (Model_1_{df}=10, $Model\_2_{df}$ =16, $\chi^2(6)$=122.92, $p < 0.0001$) despite being penalized for containing many additional parameters. This interaction makes sense, as my history of Terran experience will be most relevant to games where I am playing as Terran.

In summary, it appears that the presumptive definition of experience is problematic in StarCraft 2.[16] Models are improved by ascribing individuals with race-specific learning slopes. This suggests that different races learn at different rates.

It may also suggest that experience with specific races does not transfer perfectly, at least insofar as perfect transfer implies that experience in one race should have the same effect as experience in another race. According to $Model_3^{Pr,Tr}$, for example, a Terran player with only 100 games of experience in Terran would have a different median FAL from a first-time Terran player with only 100 games of experience in Protoss. The presumptive model, on the other hand, would not have different predictions in these two cases. The success of modelling interactions between race-specific experience and current race in $Model_3^{Pr,Tr,Interaction}$ suggest that races not only learn at different rates (e.g., 100 games worth of Zerg may be associated with a larger speed increase than 100 games of Terran), but also that the relevance of my Zerg experience to a given model prediction will vary depending on the race

[14]I will not add Ze_{xp} to the model will as $1v1_{xp} = Pr_{xp} + Tr_{xp} + Ze_{xp}$.

[15]For brevity, I include only $Model_3^{Pr,Tr,Interaction}$, which contains three terms for various interactions between race and race-specific experience, though the results are not changed if one decides to test one interaction term at a time.

[16]Given that calculating influential points is computationally demanding with this dataset, a more careful consideration of possible influential points will be delayed until the next section where I produce more interpretable models which also allow for the same conclusions.

played within the game. If complexities of skill transfer were not in play, prior experience with Protoss would not have a reduced weight if one happened to be playing Terran.

Finally, the success of these models is the first hint that researchers need to think of experience as multi-dimensional in StarCraft 2. Aggregating race experience (Te_{xp}, Ze_{xp}, and Pr_{xp}) into a single measure of experience into a single variable ($1v1_{xp}$), appears to have resulted in a loss of predictive utility.

3.6 Restricting Analysis to Dominant-Race: Probing Transfer Effects

The previous section has shown that race-specific experience does not exhibit perfect transfer to other races (because if transfer was perfect then $1v1_{xp}$ would be a perfectly respectable operationalization of experience). Unfortunately, the models in the previous section are poor tools for judging whether cross-race skill transfer for three reasons. First, $Model_3^{Pr,Tr,Interaction}$ contains highly correlated experience predictors ($1v1_{xp}$ and Tr_{xp} exhibited a Pearson's r of .8 and $1v1_{xp}$ and Pr_{xp} had a Pearson's r of .4), making interpretation of coefficients highly problematic. Secondly, the model is highly complex, making interpretation awkward even if the parameter estimates could be trusted. For a summary of this model, and for a clearer understanding of its complexity, see the Appendix ii.1.

Finally, it remains *possible* that the success of $Model_3^{Pr,Tr,Interaction}$ was due to a total lack of transfer. The importance of race-specific experience measures might have been due to the fact that about 4.6% of the games were off-race (games where a participant is not playing as their dominant-race). If one is explaining the performance of someone who usually plays Terran and sometimes plays Protoss, for example, then indicators of Protoss experience might be especially valuable when explaining Protoss performance.

If one's goal is also to evaluate whether cross-race transfer *exists*, and ideally estimate its impact on performance, then one will need an analysis which is targeted at specific kinds of transfer. For example, the data contain very few off-race games and so I will probably be unable to evaluate the extent to off-race Terran and Protoss experience transfer between one another in a participant who uses Terran as a dominant-race. A more realistic strategy is to restrict the analysis to the explanation of performance in Dominant-race games. I can then use experience with the dominant-race (Dom_{xp}) as my primary indicator of experience, and any effect of off-race experience (Off_{xp}) would be seen as a source of transfer. The order of nested models I consider remains the same.

1. $Model_1^{Dom} : FAL \sim Dom_{xp} + (1|Participant_{ID})$
2. $Model_2^{Dom} : FAL \sim Dom_{xp} + Race + (1|Participant_{ID})$

3. $Model_3^{Dom} : FAL \sim Dom_{xp} + Race + (1 + Dom_{xp}|Participant_{ID})$[17]

4. $Model_4^{Dom} : FAL \sim Dom_{xp} + Race + Off_{xp} + (1 + Dom_{xp}|Participant_{ID})$

Dropping games that were played off-race resulted in a dataset of 78,936 games and 103 players. Importantly, Figure 3.12 reveals that only 18 players have off-race games, and most of these belong to the most extreme players. I therefore need to be especially wary about generalizing any of the findings about cross-race transfer to future datasets.

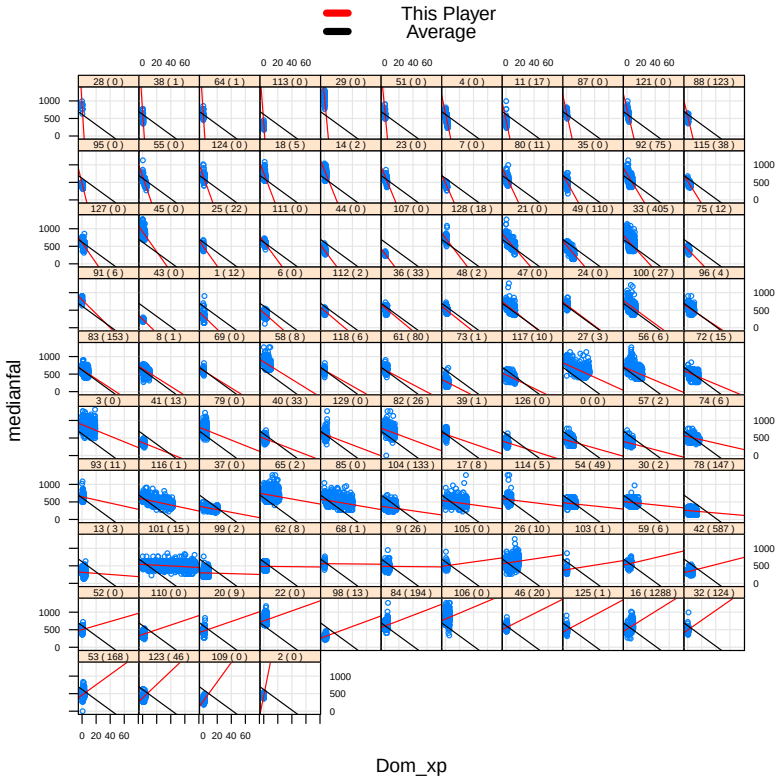


Figure 3.11: Lattice plot of Dominant Experience against median FALs, ordered by a simple regression line for the individuals data. The number of off-race games are reported in parentheses). The black average regression line reflects the fixed intercept and slope from the best model, $Model_5^{Dom}$.

$Model_4^{Dom}$ was the most successful of the analysis of nested models (all likelihood ratio tests are described in Appendix ii.2). $Model_2^{Dom}$ beat model $Model_1^{Dom}$ (Model_1_{df}=4, $Model\_2_{df}$ =6, $\chi^2(2)$=10.43, $p = 0.0054$), $Model_3^{Dom}$ beat model $Model_2^{Dom}$ (Model_1_{df}=6, $Model\_2_{df}$ =8, $\chi^2(2)$=13906.12, $p < 0.0001$), and $Model_4^{Dom}$ beat $Model_3^{Dom}$ (Model_1_{df}=8, $Model\_2_{df}$ =9, $\chi^2(1)$=4.46, $p = 0.0348$).

[17]For a summary of the presumptive model using dominant-race data, see Appendix ii.2.

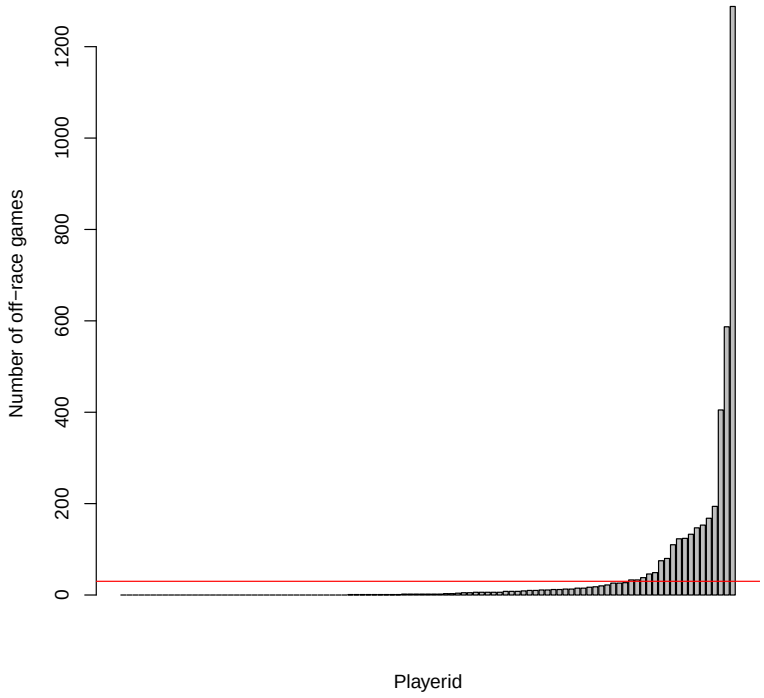


Figure 3.12: Barplot of the number of off-race games by player. the red line corresponds to 30 games.

3.7 Is the New Presumptive Model Overly Restrictive?

I have decided that the original operationalization of experience is overly broad in the sense that it glosses over race-specific experience. I can now ask whether the focus on race-specific experience is also overly restrictive in the sense that it ignores experience in other StarCraft 2 game types, such as 2v2, 3v3, and 4v4 games.

I have a relatively small number of games with more than two players in the sample, so the analysis will necessarily be simplified by focusing on 2v2 games (there are 7,729 meaningful 2v2 games in the database) and competitive games which are neither 1v1 nor 2v2 games (of which 2,239 remain in the sample). As explained in Chapter Two, my primary interest is in explaining 1v1 median FALs, so I do not add these games to the dataset. Instead, I use the presence of these games to keep track of the prior experience of players when they are participating in 1v1 games. My research question will be whether I can improve the models of median FAL by adding these other forms of experience to the model.

$$Model_5^{Dom}: FAL \sim Dom_{xp}+Race+Off_{xp}+2v2_{xp}+(1+Dom_{xp}|Participant_{ID})$$

$Model_5^{Dom}$ beats $Model_4^{Dom}$ (Model_1_{df}=9, $Model\_2_{df}$ =10, $\chi^2(1)$=231.17, $p < 0.0001$), suggesting that 2v2 xp does transfer into 1v1 performance speeds.

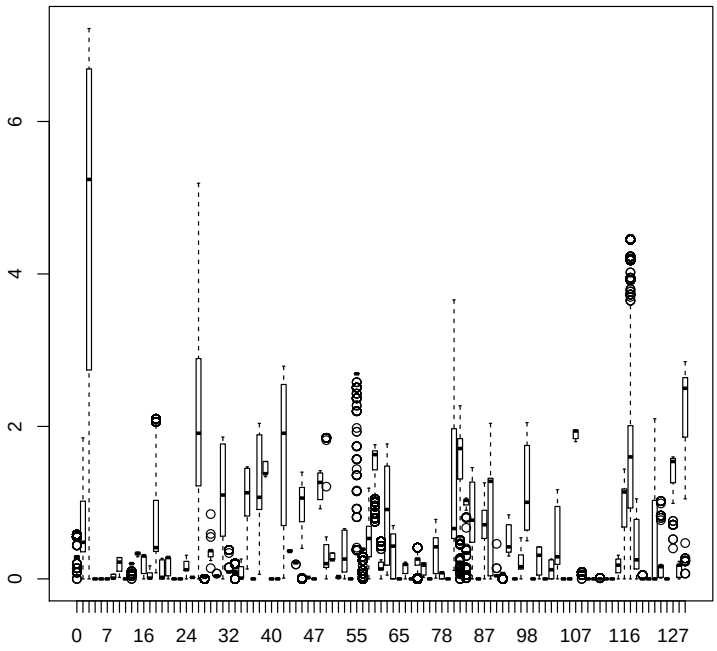


Figure 3.13: Boxplot of 2v2 xp by participant ID.

Finally, I consider whether other kinds of games (which I call N-2v2 games) also transfer.

$$Model_6^{Dom}: FAL \sim Dom_{xp}+Race+Off_{xp}+2v2_{xp}+N-2v2_{xp}+(1+Dom_{xp}|Participant_{ID})$$

$Model_6^{Dom}$ fails to beat out $Model_5^{Dom}$ (Model_1_{df}=10, $Model\_2_{df}$ =11, $\chi^2(1)$=1.32, $p = 0.2510$). However, it is worth noting that the order in which I add terms to the model does matter here. $Model_6^{Dom}$ would beat $Model_4^{Dom}$ given the opportunity. Dropping points resulting in a very high leverage in any model of median FAL in dominant-race performance will not impact model performance. If one drops the 4,917 games with a leverage value more than six times the mean leverage value, they will retain the same pattern of results and similar parameters.[18]

I use Leverage rather than Cook's Distance here to identify influential points because of run-time issues in calculating Cook's Distance over the entire dataset. I can, however, employ Cook's Distance to calculate whether some *players* are unduly influential. If I calculate

[18]If one applies more stringent criteria, such as the criterion that a game has three times the mean leverage (Pardoe, n.d.), one drops 19,406 data points (about 25% of the dataset) and, not surprisingly, gets quite different results. The factors *Race* and Off_{xp} do not yield better models in these cases, while $N - 2v2_{xp}^{Dom}$ does. Obviously parameter values shift considerably. I don't take those results too seriously here, however, as this requires dropping almost a quarter of the data as having undue influence on the model.

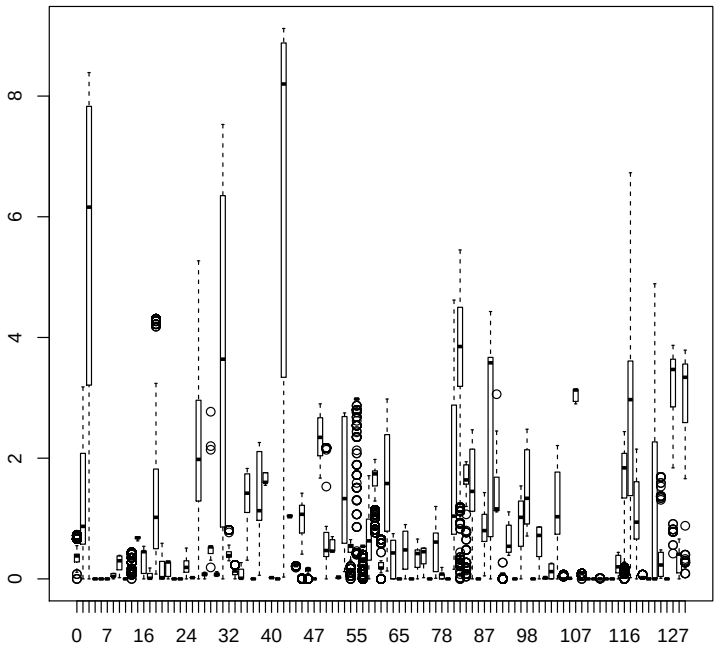


Figure 3.14: Boxplot of N-2v2 xp by participant ID.

Cook's Distance based on the exclusion of individuals players, then five influential players stand out as having a Cook's Distance greater than three times the mean (*Cook's Distance*, n.d.). Following conventional rule of thumb, I consider dropping these individuals (Lane, n.d.). Rerunning the analysis with the remaining nine players and 69,470 games result in the same pattern of results, with $Model_5^{Dom}$ winning out.[19]

In either case, it seems clear that 2v2 experience does impact single player speed. I probe the nature of this impact in the next section.

3.8 Evaluating the Final Model

Residuals for $Model_5^{Dom}$ are reasonable and shown in Figure 3.15 (for a summary of the model, see Appendix ii.2). As before, I need to be careful about making claims regarding median FALs that are over one second, as I lack much data on these slow entities.

Broader learning patterns can be examined using the best model in Figure 3.11. I do not use standardized data here, as examining learning carefully requires contemplating both the slope and intercept of individuals. For example, it is common in longitudinal data to

[19]Importantly, this drop does impact parameter estimates as I will discuss in later sections.

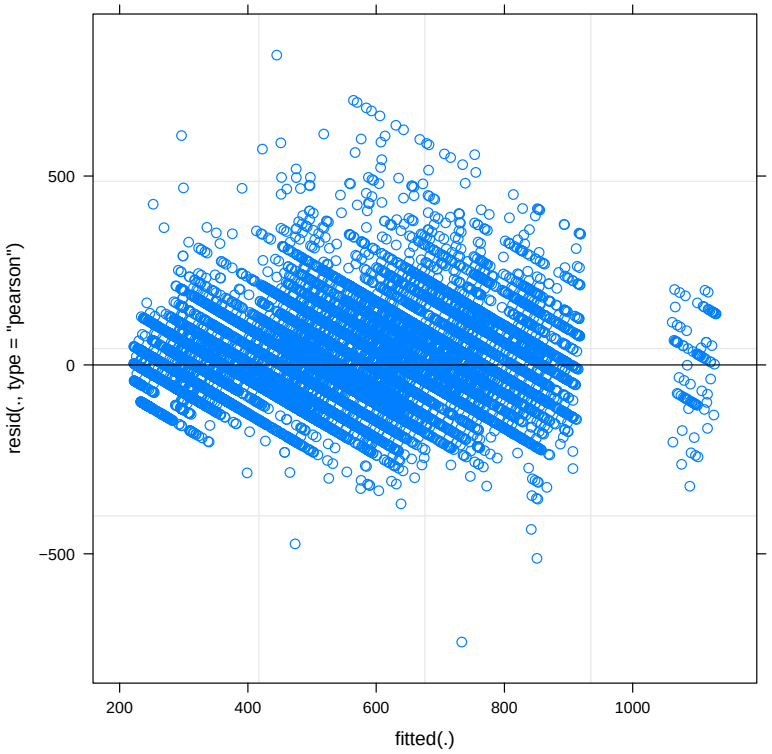


Figure 3.15: $Model_5^{Dom}$ Residual plot.

find that negative slopes have high intercepts and that positive slopes have relatively low intercepts. This can be explained by regression to the mean because (Barnett, Pols, & Dobson, 2005), even if no player is learning, some will have more extreme intercepts than others (simply because they happened to start out with a few strong or weak games). This can result in artifactual slopes. This might be somewhat at play in Figure 3.11, as there seem to be a lot of higher intercepts associated with the most negative individual slopes (see top row, red regression lines), and lots of low intercepts associated with the most positive slopes (bottom row, red regression lines).

A second possible issue is that starting strong leaves a player with less room to improve, so faster intercepts result in less impressive learning curves. Figure 3.11 does show that there are a number of players with faster than average intercepts who nevertheless show learning (e.g., see rows five and six). However, these players do tend to exhibit weaker learning effects than the players with slow intercepts (e.g., first two rows).

Despite these methodological cautions, Figure 3.11 is suggestive of learning effects. Despite a wide variety of intercepts for individuals (see red regression lines), I see negative relationships as far as the seventh row (e.g., player 17). To get a clearer sense of the learning reflected in the data I turn to profile plots examining how model quality falls off with the manipulation of fixed effects parameters. Figure 3.16 shows that dominant-race experience,

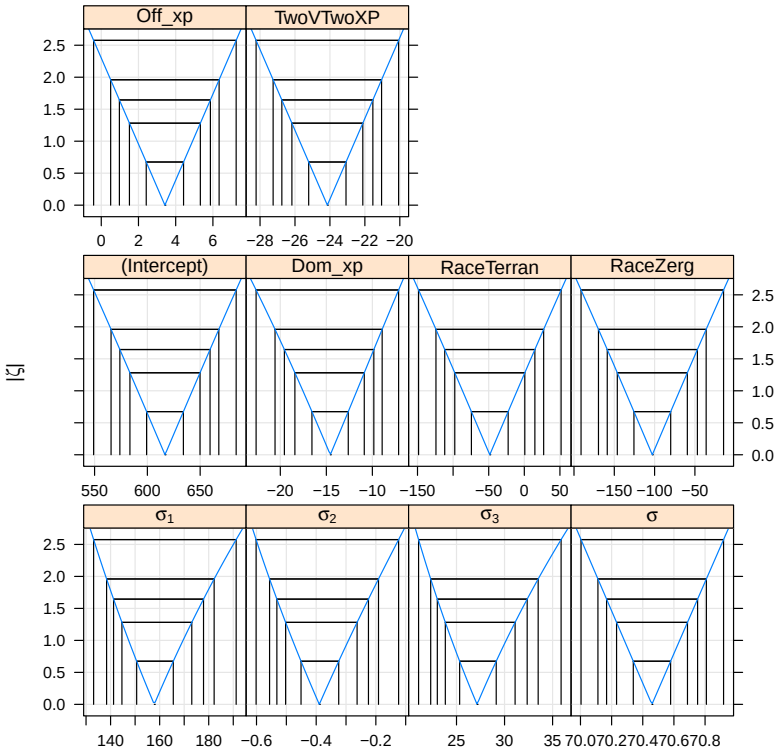


Figure 3.16: Profile plot of the best model, $Model_5^{Dom}$. Boundaries reflect (from outside to inside) 99%, 95%, 90%, 80%, and 50% confidence intervals respectively. σ refers to the standard deviation of residuals, σ_1 to the standard deviation of the random intercept, σ_3 to the standard deviation of the random slope, and σ_2 , to the correlation between the random slope and intercept.

on average, is associated with a 14.6 millisecond decrease every one hundred games (95% CI [8, 22 milliseconds]).

One hundred games of off-Race experience appear to yield a slight slowing of 3.5 milliseconds positive to median FALs (95% CI [0.5,7], see Figure 3.16). I also examine the impact of off-race experience in comparison to dominant-race experience using what I call a *transfer coefficient*[20] $\frac{Off-Race_{xp}}{Dom-Race_{xp}} = -0.23$. It appears, therefore, that off-race experience disrupts Dominant-race performance by about 23% of the extent to which a Dominant-race game improves performance. This is not surprising given that players with many off-race games exhibit some of the most slowing in Figure 3.11 (The number of off-race games are shown in parentheses). It also fits with the general finding that expertise is extremely domain specific.

It is worth noting, however, that these parameter estimates are sensitive to a small set of influential players. If I drop influential players with a Cook's Distance of more than

[20]Thanks goes to Jack Davis for providing me with the idea of using a transfer coefficient.

three times the mean (*Cook's Distance*, n.d.), parameters for off-race shift somewhat. For example, excluding 9 influential players whose Cook's Distance was more than three times the mean resulted in the same optimal model structure $Model_5^{Dom}$. This model had similar parameter estimates except off-race experience dropped to -9. Across these 9 players and 69,470 games, off-race experience seemed to be a positive influence, worth about 66% of the speed gains associated with a 1v1 match. I do not take this model too seriously given that it requires dropping such a large portion of the dataset, but it does seem likely that individuals differ in terms of how off-race experience impacts their performance.[21] I now turn to a more careful interpretation of the $2v2_{xp}$ parameter, which requires contemplating the existence of third factors which could impact my results.

3.8.1 Interpreting the Magnitude of Team-Game Transfer Effects: Ruling out Dedicated Practice as the Potential Source of a Simpson's paradox

In the best model, $2v2_{xp}$ is surprisingly associated with declines of 24 milliseconds per 100 games (95% CI [21,27]; Transfer coefficient=1.65, see Figure 3.16). It is plausible that 2v2 games would sometimes lead to improvement, as they allow players to observe another's performance directly, opening up the possibility of observational learning effects. However, I need to be very cautious about inferring that 2v2 games have 1.65 times the value of a 1v1 game. One reason is the relatively small number of these games in the dataset. However, there is a deeper reason I should be cautious about making this interpretation.

Ignoring possible third factors can lead to a variety of misleading results in the aggregate level (Simpson, 2017). As a simple example, consider if two variables are positively correlated on one level of a factor but negatively correlated on another. The variables may appear to be unrelated at the aggregate level (Kendall, 1945 as cited in Simpson, 2017). Similarly, there are cases where one can find entirely different correlations depending on whether data are aggregated together or separated by group.

One factor that might impact the results is that people who play 2v2 games might choose to do so because they become bored with typical StarCraft 2, while players who stick to 1v1 games might be more dedicated. Since dedicated individuals are more likely to practice with the intention to improve, and since this sort of deliberate practice is thought to be crucial to expertise (Ericsson et al., 1993), it is plausible that players with more 2v2 games have more to learn. Since novices (with slow median FAL intercepts) are likely to have more impressive slopes than experienced individuals, the apparent utility of $2v2_{xp}$ may be an artifact of including both dedicated and non-dedicated players in the analysis. Previously, this possibility seemed somewhat unlikely given Figure 3.17, which suggests that

[21] One gets similar results if one instead uses other rules of thumb for dropping influential points, such as excluding players with a Cook's Distance greater than 1 (Lane, n.d.).

while there is a general lack of $2v2_{xp}$ in this dataset, there do seem to be a collection of relatively fast and slow games played by players with over 50 2v2 games.

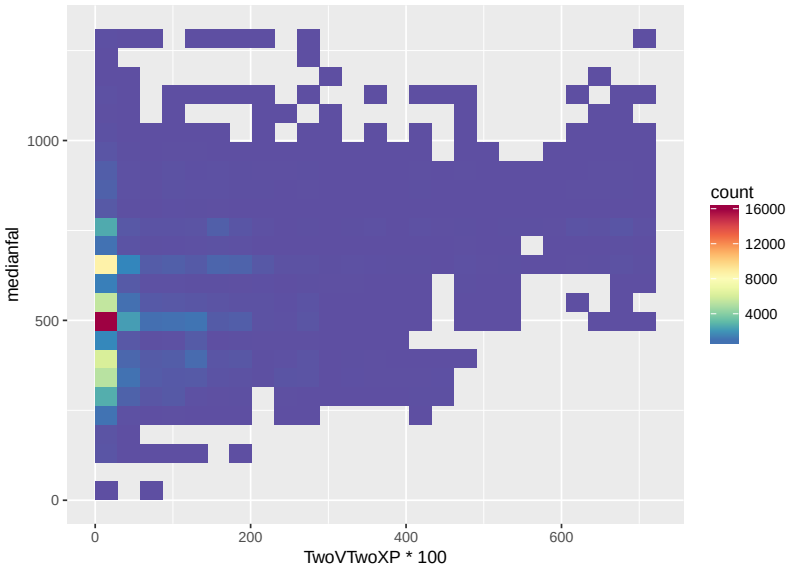


Figure 3.17: Plot of median FAL against 2v2 experience. Colours reflect data frequency. Note that data are not independent (i.e., each participant contributes multiple observations).

Rather than attempt to measure player dedication directly, I began by simply restricting the analysis to players with 30 or more 2v2 games. Highly dedicated players who restrict themselves to 1v1 games will therefore not impact this analysis. This required dropping 74 players, resulting in a dataset of 29 players and 23,451 games. The best model emerging from this analysis was $medianFAL \sim Dom_xp + Race + 2v2_{xp} + (1 + Dom_xp|Participant_{ID})$. The coefficients for Dom_{xp} and $2v2_{xp}$ were -14.7 and -23.15 respectively. In other words, even in a model of players who have a history of $2v2_{xp}$, dominant-race experience is associated with smaller reductions in action latency than multiplayer experience.[22]

It seems unlikely that the seemingly exaggerated value of $2v2_{xp}$ games is due to better players avoiding this mode of gameplay. Restricting the dataset to players who do sometimes

[22]Notice that this particular problem could conceivably apply to the consideration of off-race experience as well. If players who play off-race are less dedicated, then I would expect them to have more impressive learning effects than those who stick to their dominant-race. In order to rule out the possibility that my estimates of off-race experience underestimated the disruption caused by playing off-race, I reran the analysis after dropping 85 players for having less than 30 off-race games, resulting in a dataset of 12,848 games across 18 players. The best model emerging from this analysis was $medianFAL\ Dom_xp + Off_xp + (1 + Dom_xp|Participant_{ID})$, the coefficients for Dom_{xp} and Off_{xp} were -8.8 and 3.3 respectively, implying that an off-race game disrupted learning by about 40% of what a Dominant game contributes to learning. This suggests that the original model, with a transfer coefficient of 23% may have been influenced somewhat by players with a few off-race games. Further, $2v2_{xp}$ did not improve model performance in accounting for these data, even though these 18 players shared 1,560 2v2 games between them.

play 2v2 games did not eliminate the result. Furthermore, Figure 3.17 suggests that there are some relatively fast players with more than 50 games of $2v2_{xp}$.

One limitation of my previous attempts to resolve the puzzle is that I did not measure the dedication of players directly. Instead I only assumed a relationship between dedication and the number of 2v2 games played. In order to attempt a direct measure at dedication, I turned to six Likert-scale survey questions (never, rarely, sometimes, often, and very-often) about player training schedules.

> Do you use any of the following training strategies?
>
> 1. Play on special skills/practice maps
> 2. Look up strategies online
> 3. Watch your own replays
> 4. How often do you play ladder games with the intent to improve a skill rather than solely for the purpose of winning the game?
> 5. Watch professionally narrated streamed games
> 6. Participate in online communities to talk about SC2

Question 1 to 4 are intended to gauge a player's interest in practicing to improve (Ericsson et al., 1993) rather than to simply play for leisure. The maps discussed in question 1 are usually intended for improvement rather than enjoyment, question 2 probes whether participants pursue background knowledge related to the game, question 3 could reflect deliberate practice insofar as players watch their old games in order to improve, and four directly inquires about deliberate practice. Question 5 and 6 speak to indirect experience with 1v1 StarCraft 2, although they do not necessarily imply deliberate practice. Players who watch professional StarCraft 2, or discuss StarCraft 2 online presumably have additional knowledge about 1v1 gameplay. This is because professional StarCraft 2 is almost exclusively made up of 1v1 games, and discussions around skill in StarCraft 2 are usually discussions of 1v1 performance.

I scored individuals by taking the number of training strategies they employ 'often' or 'very-often'. However, I take this as a gross measure and do not provide evidence that it is a reliable or valid measure of motivation to improve.[23] Indeed, a look at Figure 3.18 and Figure 3.19 do not even suggest that more motivated players are necessarily better (Spearman's

[23]Of course, the survey does have other limitations which are worth considering. First, the survey questions may be susceptible to bias. Given that the purpose of the study was transparent, and given that the training strategies discussed in the questions are familiar to StarCraft 2 players, it may be that some weaker players were impacted by a social desirability bias and exaggerated their use of training methods.

Secondly, the scoring of responses to survey questions emphasized convenient interpretation (i.e., the number of training strategies that players use often) despite the potential loss of information that comes with effectively dichotomizing six Likert responses into 'Often' versus 'Not-Often'.

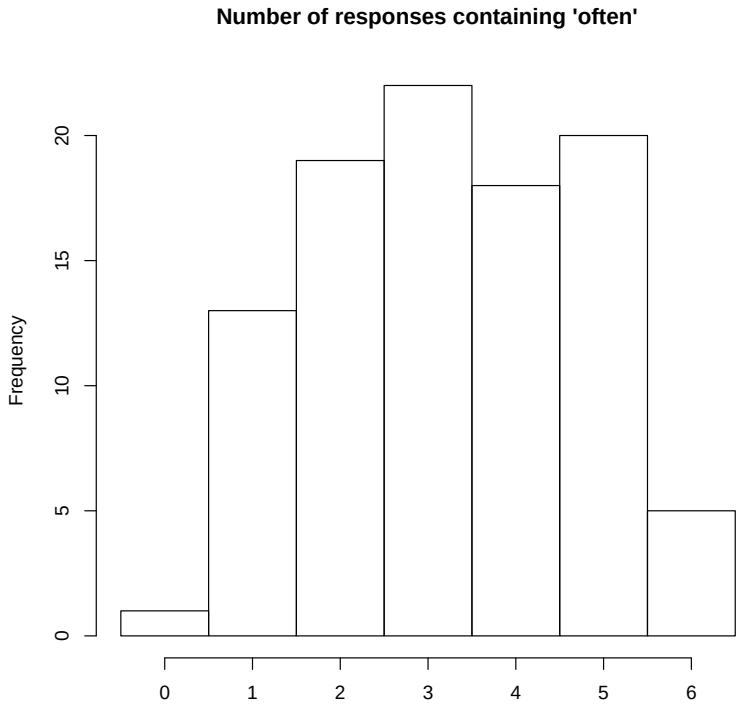


Figure 3.18: The number of times a participant responded to a training strategy question with an option containing the word 'often'.

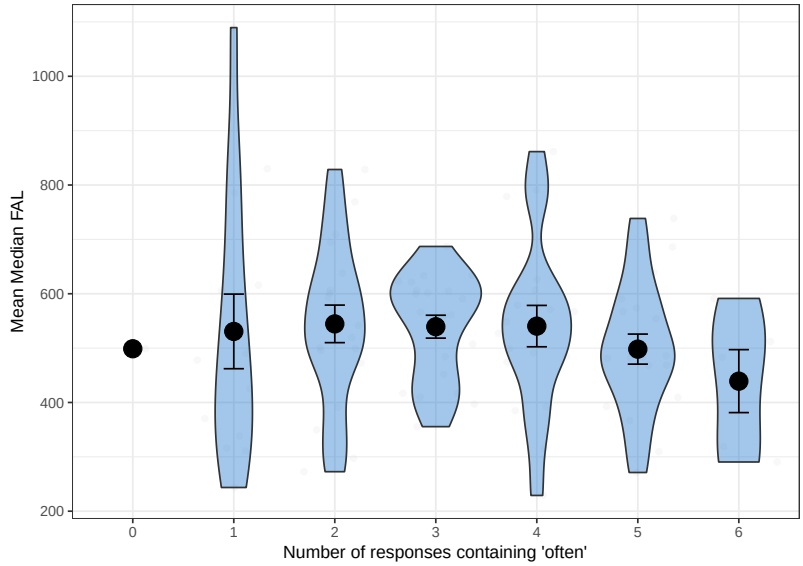


Figure 3.19: Mean median FAL by the number of times a participant responded to a training strategy question with an option containing the word 'often'. Blue fields represent a kernel density estimate.

r= -.08). This might initially seem somewhat surprising given the central importance of deliberate practice to our understanding of expertise (Ericsson et al., 1993).

Importantly, results described in Figure 3.19 do not cast doubt on whether First Action Latency is a marker of player skill. Furthermore, it does not cast doubt on the use of median FAL to characterize player skill in the present sample (e.g., Figure 3.3). Instead, I will argue that these results should be taken to mean that what was initially intended to measure 'dedication' should be thought of as an indicator of general motivation and not necessarily skill or motivation to improve at 1v1 gameplay specifically. Since motivation to improve is a possible factor which could conceivably impact the value of 2v2 experience to 1v1 performance, we still use these data here.

Thinking about dedication as a measure of motivation

There are two reasons that we should not be anxious about the failure to find a strong relationship between dedication and skill.

First, it is important to recall that the survey questions that are the foundation for my measure of dedication are quite gross. While deliberate practice theory (Ericsson et al., 1993) does predict that skill should be related to *the number of hours of deliberate practice*, it is unclear whether the fact that a player reports using deliberate practice is enough for us to make predictions about their skill. Furthermore, the longitudinal data described here samples from only 103 players while Thompson et al. (2013) studied the relationship between mean FAL and skill in a sample of 3,395 players.

Secondly, recall that only a small number of participants are anywhere close to the speeds expected of professional level players (see Figure 3.2). Based on my familiarity with the task domain, I would not find it surprising if players could reach the Platinum level with a mediocre motivation to improve and little or no deliberate practice.

I will argue that we should consider thinking of dedication as a measure of motivation rather than skill. Motivation to improve, while a necessary condition for the highest levels of performance, is likely to change with experience as casual players become bored. If so, it would not be surprising to find highly motivated Bronze-level players and unmotivated Diamond-level players. We would only expect relations between motivation and skill after carefully controlling for experience.[24]

There is also reason to think that my measure of dedication does not track 'dedication to 1v1 gameplay.' There was no evidence that dedicated players played fewer 2v2 games (Spearman's r=0.025). One possible explanation for this is that players prefer 2v2 games

[24]Such analyses are worth considering but are beyond our present scope given that dedication is only used the basis for a follow-up exploratory analysis.

because it allows them to socialize with friends online.[25] According to this explanation, the motivation to improve at 1v1 gameplay is not enough to overwhelm the desire to occasionally play with friends. Another possible explanation is that the survey is a better indicator of a players overall motivation to improve, not their foveal dedication to 1v1 gameplay.[26] On either explanation, dedication does not entail a focus on 1v1 gameplay. Instead, it only tracks an overall motivation to improve.

Results of the dedication followup analysis

I proceed with the intended analyses of dedication with the caveat that the measure of dedication is more likely to measure motivation than dedication to 1v1 play. My objective is to better understand the exaggerated value of $2v2_{xp}$, and my hypothesis is that 2v2 play means something very different to non-dedicated players. I specify players as non-dedicated if they answered 'often' or very-often' to fewer than three of the six questions. On this restriction, I am left with 33 players and 28,660 games for the dominant-race analysis. The best model, according to my series of likelihood ratio tests, was $Model_6^{Dom}$: $FAL \sim Dom_{xp}+Race+Off_{xp}+2v2_{xp}+N-2v2_{xp}+(1+Dom_{xp}|Participant_{ID})$. The transfer coefficients for $2v2_{xp}$ persisted among non-dedicated players ($Dom_{xp} = -14.24$, $2v2_{xp} = -24.12$, Transfer Coefficient=1.7). What was surprising, however, was that $offrace_{xp}$ appeared beneficial for these players ($offrace_{xp} = -25.46$, Transfer coefficient=1.91). Rather than disrupting learning, off-race experience appears to carry about 180% of the value of dominant-race experience for non-dedicated players.[27] Unfortunately, I lack the sample size to analyze $2v2_{xp}$ in dedicated players, as only two dedicated players have more than 30 2v2 games.

In short, followup analyses have revealed that off-race may be more valuable to less dedicated players who are not pursuing deliberate practice, perhaps because dedicated players learn about the other game species from other sources. I interpret this result with some

[25]In competitive StarCraft 2, the game manufacturer is responsible for finding opponents. The only way to play competitively with a friend is to play team games such as 2v2 StarCraft 2.

[26]While professional players play 1v1 StarCraft 2 exclusively, there are still ranked team matches. This makes it possible that some players have an interest in improving at both 1v1 and 2v2 gameplay while others do not.

[27]An alternative procedure would be to avoid dropping data based on dedication and instead use dedication as a variable in the analysis. I do not report this analysis here because my primary interest is in model coefficients, which become more difficult to interpret if I add necessary interaction terms between dedication and hypothesized $2v2_{xp}$. Nevertheless, adopting such an approach leads to the same effects discussed here. I produced a more complex model, '$medianFAL \sim Dom_xp + Race + Off_xp + 2v2_{xp} + dedication + Off_{xp} : dedication + (1 + Dom_{xp}|Participant_{ID})$', which was superior to the best dominant-race model (given the likelihood ratio test). This complex model involved dedication and interaction terms between dedication and off-Race experience. Interaction terms for $2v2_{xp}$ and dedication did not significantly improve model performance, and the coefficients for $2v2_{xp}$ remained similar. As expected given the previous analysis of non-dedicated players, parameter estimates of interaction terms suggested that off-race experience was more beneficial for non-dedicated players.

caution, however, as the goal of this ad hoc analysis was to explain the surprisingly strong impact of $2v2_{xp}$ on model predictions, not off-race experience.

Unfortunately, I still lack any explanation for the extreme impact of $2v2_{xp}$ to 1v1 gameplay on performance. The result appears surprising in light of the extreme domain specificity of expertise (see Chapter 1.05). More work needs to do be done to consider extraneous variables which might rule out possible artifacts. This work must, unfortunately, be ad hoc and exploratory, as our understanding of any complex skill is too underdeveloped for me to be sure that this strong learning effect of $2v2_{xp}$ is not brought about by ignoring some third factor. It is worth emphasizing that this result is not inconsistent with the domain specificity of expertise. My models may actually provide reasonable fixed effects estimates of the value of 2v2 experience in light of the fact that the players *who choose* to play 2v2 games are just the sort of people who benefit from the experience. These results do not compel me to believe that 2v2 experience would have similar benefits for players who focus on 1v1 gameplay.

3.8.2 Conclusion

It is obvious from the present work that longitudinal studies of the development of complex skills in situ should take a multidimensional approach. First, domain performance may be impacted by nuisance factors, much as how StarCraft 2 is impacted by race. Secondly, and perhaps more importantly, extraneous factors need to be considered when aggregating across a number of learning curves in order to avoid paradoxical results. While future work should provide a more careful validation of the survey data, there is some reason to think that off-race experience is more valuable to players who are less motivated. There is even reason for thinking that the present work did not consider a sufficient number of extraneous variables, as I am currently are unable to explain the value of 2v2 experience to overall performance, even if I restrict the analysis to players who occasionally play 2v2 games, or to players who appear less motivated to improve.

There is also a second, and potentially more important, respect in which this study encourages a multidimensional approach towards longitudinal studies of complex skill. The present results support the view that experience itself should be thought of as multidimensional. Intuitive definitions of experience (such as 1v1 performance regardless of race) can easily aggregate different forms of experience that contribute to performance differently. In the case of StarCraft 2, off-race experience appears to disrupt learning, at least for some classes of players. Furthermore, while the magnitude of effect estimates for 2v2 experience deserve scrutiny, it is true that model performance is enhanced by including 2v2 experience. Just as basketball players appear to have nested expertise for foul-line shots within their more general expertise for basketball play (Keetch et al., 2008), very different sorts of experience seem to play into overall StarCraft 2 performance. Where multidimensional def-

initions of experience are possible, as in digital archives of performance, researchers should consider thinking of learning curves as learning hyper-planes.

Finally, it is worth considering whether the entire expertise continuum should be represented in a single linear mixed effects model as has been attempted here. These results provide good evidence that extra-domain experience (e.g., off-race play or 2v2 experience) can have unexpected transfer effects to domain performance, and so I have made progress towards the goal of identifying a better operationalization of experience in StarCraft 2. However, it seems that the only way I can believe the parameter estimates in the best models is if I postulate that the more serious players sometimes avoid certain kinds of experiences (e.g., off-race play or 2v2 experience), making it hard to judge whether these experiences are of value to this class of player. This leads me to the notion that segments of the expertise continuum might be best understood separately, as when Thompson et al. (2013) asked whether the variables which are relevant to expertise vary with the levels of expertise being considered. Such considerations lead me to suggest that, at least in the short term, researchers should define experience on a case-by-case basis where the definition is relative both to the skill level and practice habits of the sample.

While the present chapter has implications for how we should think about and operationalize experience in a complex domain, it does not constitute a test of any specific theory advanced by the literature. Unidimensional operationalizations of experience are common, but they do not seem theoretically motivated. The following Chapter will present a case where naturalistic telemetry is used to examine predictions from a specific psychological theory about age-related differences.

Chapter 4

Finger Tapping in StarCraft 2

While age related decline in expertise is relatively well studied, particularly in typing (Bosman, 1993; Salthouse, 1984), aviation (Morrow, Leirer, & Altieri, 1992), and Chess (Moxley & Charness, 2013), researchers have only recently begun studying age-related differences *in situ* using digital records of performance. One attempt comes from Thompson et al. (2014), who found age-related declines in mean First Action Latencies after 24 years of age. Another attempt comes from Tekofsky, Spronck, Goudbeek, Plaat, and Van Den Herik (2015), who studied aging in the first-person shooting action game Battlefield 2. They obtained similar results, with optimal performance around 20 years of age, which is in keeping with Salthouse's (2009) conclusion that aging begins in early adulthood. One difficulty with integrating these sorts of studies is that they rely on very different measures. Thompson et al. (2014) focused on performance speed while Tekofsky et al. (2015) focused on high-level meta-data such as the number of kills to deaths ratio. To help better understand these effects I will conduct followup analyses on the original 2014 study from Thompson, Blair, & Henrey.

One explanation for a host of age-related differences has been the postulation of a single underlying ability which declines with age. This entity could consequently lead to observations of age-related differences in many different measures of performance. Salthouse, for example, has argued in favour of processing speed theory, the view that the single construct of processing speed is responsible for many of the observed relationships with age (Salthouse, 1996).

In the present work I will attempt to apply *processing speed theory* to the explanation of age-related differences in StarCraft 2. I do so by creating a novel measure of performance speed, Right-Click Latency (RCLatency), which is meant as an analogue of finger tapping speed, another measure which appears to decline with age (Lezak, 2004). I then conduct a series of analyses designed to test whether RCLatency can explain the age-related differences in mean FAL observed in Thompson et al. (2014). The direct contribution of this work will be a judgment about whether the age-related differences observed in Thompson et al. (2014) should be explained as a decline in clicking speed or in terms of cognitively more

complex abilities (such as decision making). It will also speak to theories about the source of laboratory age-related differences, as the choice in RCLatency is inspired by Salthouse's work (Salthouse, 2000).[1]

There are a few complications which need to be addressed before I can proceed with my analyses. First, I take the present chapter to be an honest attempt to apply the ideas in processing speed theory, and this requires some careful discussion of what *processing speed* actually means. This will also ensure that my results will speak to the theoretical utility of processing speed theory.

Secondly, the present work will employ a slightly different statistical method than previous work. Thompson et al. (2014) used segmented regression models to estimate when the effect of aging begins, while I will follow the recommendations of Muggeo (2008), who has developed an R package that is specifically designed to overcome common pitfalls in segmented analyses. Thompson et al. (2014) used likelihood ratio test statistics to test whether adding two additional parameters, one estimating the onset of a new-age related change and one estimating the strength of this new aging effect, results in a better model. Unfortunately, the null distribution corresponding to these Likelihood Ratio Statistics is not obvious in segmented models (Muggeo, 2016). Consequently, I will rerun the original analysis using the R package 'segmented' (Muggeo, 2008), and consider whether the results are consistent with previous work (Thompson et al., 2014).

Thirdly, I need to develop the measure of RCLatency. A considered definition requires a number of exploratory analyses to ensure I understand my own measure. One such analysis, for example, involves considering whether players vary the distance between their right-clicks on the screen. If so, RCLatency might not measure tapping speed as much as it measures the distance between clicks. After dealing with these three issues, I will create an analysis strategy which addresses the question of whether RCLatency can explain age-related differences in mean FAL.

4.0.1 What is Meant by Processing Speed

Salthouse has held onto the same basic account over the years. In 2000, Salthouse rephrased this account in terms of general and specific influences on age-related declines in speed. Processing speed theory, of course, predicts that these declines have a shared (or general) source. The best evidence for processing speed theory comes from patterns of correlations observed in studies of age-related change. For example, addressing the source of laboratory cross-sectional and longitudinal differences in performance with aging (Fozard, Vercruyssen, Reynolds, Hancock, & Quilter, 1994; Salthouse, 1998) requires a careful consideration not only of relations between aging and performance measures, but also consideration of the

[1]Thanks should be given, however, to an unknown commentator at the 2014 Association for Psychological Science annual convention, who directed me to processing speed theory.

correlations between the performance measures themselves. The various relata of age are typically correlated with one another (Salthouse, 1996, 2017), making it hard for researchers to rule out the existence of a single source of age-related declines.

In order to clarify what is meant by processing speed, I begin by considering the operationalizations and methods employed by Salthouse. This will provide hints as to what is meant by a general processing decline. I will sidestep philosophical debates as much as possible, except where I think some philosophy allows for a clearer and more charitable description of processing speed theory.

Processing speed measures must, on the one hand, be suitably general that they are not tracking specialized processes such as skill in basketball or even language. On the other hand, there is some reason to think that a processing speed measure should be a measure of *cognitive processing*. '[...The] speed measure should not merely represent input and output processes or sensory and motor processes, or else it may not reflect the duration of relevant cognitive operations' (Salthouse, 1996, p. 407). One commonly used measure is the digit symbol substitution test, a task where participants must translate digits into symbols as fast as possible using a key provided by the experimenter (Salthouse, 1996, 2017). However, there is also reason to think that Salthouse would still be interested in measures of low-level motor speed. Salthouse (2000), for example, classifies finger tapping speed as an indicator of *psychomotor speed*, and proposes that such measures might exhibit age-related decline alongside a host of measures because of a postulated common cause.

Importantly, Salthouse does not believe that age-related differences observed in laboratory measures can be fully explained by differences in psychomotor speed (Salthouse, 1996). He bases this conclusion on findings of relationships between age and more cognitively interesting variables even after statistical control of motor speed. Consequently, if we develop a measure of psychomotor speed in StarCraft 2, in the form of a RCLatency, Salthouse would likely predict relationships between RCLatency and mean FALs. However, *insofar as RCLatency is a measure of psychomotor speed*, Salthouse would most likely predict that RCLatency would not fully explain the age-related differences observed in mean FAL.

Of course, this only goes some way towards clarifying what is meant by *processing speed*. One possible start is a computational theory of mind, where processes are operations over mental representations – which themselves need to be structured in such a way as to make these operations possible (e.g., see Fodor, 2008).

I would argue here that it is wrongheaded to think of processing speed theory as requiring a traditional notion of information processing in terms of operations being performed over structured mental representations (Fodor, 2008). When Salthouse is given opportunity to speculate about the cause of age-related changes across many variables, only some of the hypothesized causes are in the idiom of information processing.

> Ultimately, of course, I would like to know why increased age during the adult years is associated with decreased levels of speed, and this is another area where

> psychophysiological and neurobiological research can be expected to make important contributions. Among the speculations proposed to account for age-related slowing are that because of diffuse cell loss the transmission of neural impulses must traverse lengthier and more circuitous pathways to reach the same end state (Cerella, 1990; Salthouse, 1985), that a slower propagation of neural impulses with increased age is attributable to a reduction of dendritic branching, a decrease in the number of active synapses, or a loss of myelin (Miller, 1994), and that age-related slowing may be a consequence of a loss of synchronization of neural impulses, possibly due to a reduction of particular neurotransmitters such as dopamine. It seems unlikely that these possibilities can be distinguished with only behavioral research, and thus research with various types of psychophysiological or neurobiological variables may be necessary to help resolve the fundamental issue of the causes of age-related slowing (Salthouse, 2000, p. 25).

Importantly, there is no attempt here to spell out how cell death, a loss of myelin, or a loss of specific neurotransmitters relate to processing speed. Instead, it seems that Salthouse views processing speed theory as potentially fitting any of these explanations. Salthouse intends to let science settle the big questions about how processing speed is implemented in the brain.

> Among the important issues to be investigated are the neurophysiological basis for age-related slowing and what the processing-speed construct actually reflects (Salthouse, 1996, p. 425).

I take it that Salthouse may be under the impression that we have an imperfect understanding of what *processing speed* actually is, or what the concept means. For researchers such as Cronbach and Meehl (1955), for example, clarification of the meaning of constructs is an empirical issue. If this is the sort of view that Salthouse adopts, then he would likely not claim that the Digit-Symbol Substitution Test, or any test for that matter, could be a perfect measure of processing speed. Instead, Salthouse would think of *processing speed* as a construct that is imperfectly accessed through a variety of measures.

> Finally, as with the assessment of any theoretical construct, it is generally desirable that the construct be evaluated with several measures to minimize the specific variance associated with single measures and to emphasize the common, construct-relevant variance (Salthouse, 1996, p. 407).

Salthouse seems to be thinking of *processing speed*, therefore, as a construct that is both imperfectly understood and imperfectly measured. This vagueness in definition is expected as, according the classic account of construct validity theory (Cronbach & Meehl, 1955), the

meaning of constructs is actually clarified by the nomological network, an interconnected set of laws describing how theoretical entities relate to observables and to each other. This would explain any vagueness in the 'processing speed' construct, '[since] the meaning of theoretical constructs is set forth by stating the laws in which they occur, our incomplete knowledge of the laws of nature produces a vagueness in our constructs' (Cronbach & Meehl, 1955, p. 294). I will call this view the empirical approach to the clarification of constructs.[2]

The empirical approach towards the clarification of concepts makes it especially difficult to test Salthouse's theories in novel domains. It implies that there will necessarily be uncertainty as to whether my measures track theoretical constructs of interest. If we find, for example, that age-related variance in mean FALs can be completely explained by RCLatency, Salthouse could avoid disconfirmation of processing speed theory by insisting that RCLatency must measure more than just motor speed.[3]

I have two responses to the possibility of such ad hoc responses. First, it is useful to acknowledge that the ability of objectors to formulate ad hoc responses is not unique to naturalistic telemetry. Since construct validation, according to Cronbach and Meehl (1955), is never complete, such responses are always possible. Secondly, even if we cannot strictly confirm or disconfirm Salthouse's theory here[4], the present investigation can still be used to test whether *processing speed theory* is a useful tool for explaining age-related differences in a particular complex domain. Insofar as Salthouse's theory helps me understand observations in StarCraft 2 research, we can at least attest to its utility in explaining complex behaviour.

I begin by rerunning the original analyses with a slightly revised analysis using the *segmented* package in R (Muggeo, 2017; R Core Team, 2013). The original analysis compared piecewise models using a likelihood ratio test and a Chi-squared distribution. Instead, I follow the recomendations of Muggeo (2016), which are implemented in the *segmented* package. After rerunning the Thompson et al. (2014) analysis, I will define a new measure of speed

[2]It is important to point out that this reading of constructs is not the only account one could adopt. For the philosophical complications surrounding this understanding of 'construct' see the discussion of Slaney and Racine (2013) and Slaney (2017). Thanks goes to Kathleen Slaney for pointing out that one might prefer to treat definitional issues surrounding the concept of *processing speed* by appeal to conceptual analysis alone. Questions about the neural underpinnings of processing speed would remain empirical questions. I have no objection to such a view here, and I do not believe it would substantially change the interpretation of my results. I only give special credence to the aforementioned understanding of construct as it is a popular reading of Cronbach and Meehl (1955) which, as the reader will see, could be used as the basis for an objection to the present work.

[3]It is important to acknowledge, of course, that such responses, while ad hoc, may turn out to be useful if they at least offer up new empirical predictions.

[4]Note that, construct validity issues aside, we cannot strictly disconfirm Salthouse's theory here as this view is explicitly about a general factor which explains age-related differences in a variety of variables. Salthouse (1996) is not proposing that every measure of speed will show age-related decline because of a single common cause.

in StarCraft 2, right-click speed, which appears to track speed but which puts less emphasis on decision making.

4.0.2 Rerunning Thompson et al. (2014)

Thompson et al. (2014) used piecewise linear regression to estimate the onset of aging. These piecewise models contained a split point parameter which represents the year at which the effect of aging changes. Consequently, they also contain two aging variables, one reflecting aging and one reflecting the number of years above the split point. Various split points were then compared with likelihood ratio tests using a Chi-Squared distribution. I rerun this analysis with two changes.

First, I begin by rerunning the original initial analysis using the R package 'Segmented' (Muggeo, 2017). After building these models, I compare the segmented model, $Model^{segmented}$, against a simpler model that is not segmented.

$$Model^{linear} : meanFAL \sim (league * age)$$
$$Model^{segmented} : meanFAL \sim (league * age) + age_{segmented}$$

The models are nested, so I will compare them with a likelihood ratio test.

Secondly, I rerun the analysis using the revised parsing of the skillcraft data from Thompson, McColeman, et al. (2017), which has some minor differences from the original study (Thompson et al., 2013). This will allow me to compare the performance variables defined in the next section to First Action Latencies using the same dataset and parsing procedure. The largest difference between the two datasets is a slight difference in sample size (23 games are dropped from the newer dataset due to parsing errors).

Importantly, Thompson et al. (2014) rely on *mean* First Action Latencies rather than the medians used in the previous chapters. Since the present chapter is meant to build off of the original study, I will continue to use mean First Action Latencies in order to make direct comparison between the studies possible.

Results were in keeping with Thompson et al. (2014) (for further details of this analysis, see Appendix ii.4). The simple linear regression model $Model^{linear}$ is nested within $Model^{segmented}$, which has two additional parameters: one for the breakpoint at which the impact of aging changes, and one for how the effect of aging changes after the breakpoint. I retain both league and age as variables in the base model on theoretical grounds, as a common finding is that the impact of aging on performance is ameliorated by skill (Bosman, 1993; Salthouse, 1984). Since the model is nested, I compared the segmented model to the simpler linear regression using a likelihood ratio test, with the segmented model winning out ($Model^{linear}_{df}$=13, $\text{Model}^{segmented}_{df}$=15, $\chi^2(2)$=11.07, $p = 0.004$).

The optimal breakpoint, which represents my best guess of when age-related decline begins, remains 24 years of age. The confidence interval (based on two times the standard

error of the segmented model) is 21-27, which is somewhat tighter from the 20-29 years of age reported by the original study. As recommended by the R package 'Segmented' (Muggeo, 2008), a Davies test was used to test whether the effect of age differed before and after 24 (number of possible splits=27; p=0.012).

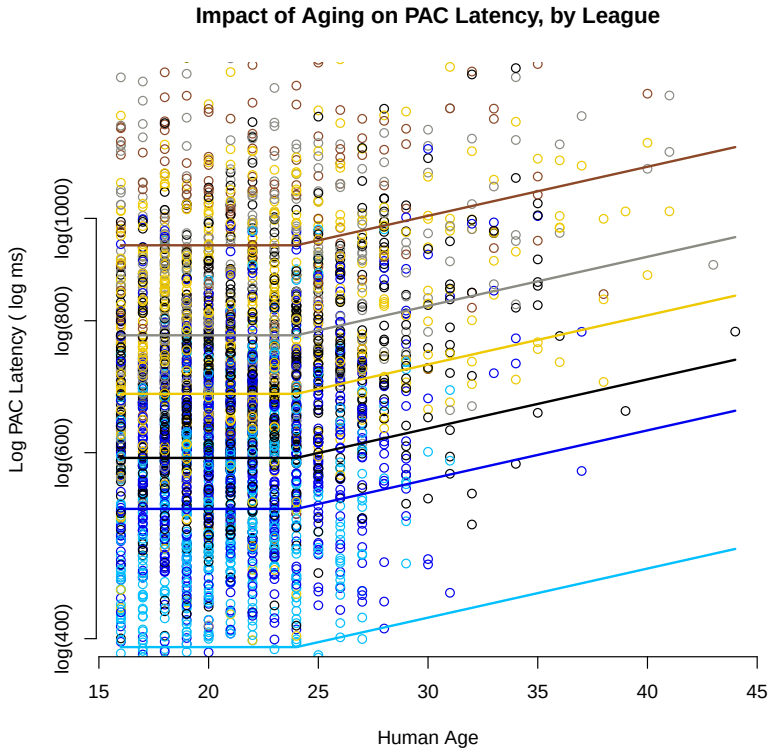


Figure 4.1: Segmented model of mean FAL on age and league. Estimates of interactions between mean FAL and league are not reported as no interaction was significant at a familywise error rate of 0.1. The differences in the intercept between Silver and Bronze, while depicted in the figure, are not statistically significant. Coloured lines (from top to bottom), represent 'Bronze', 'Silver', 'Gold', 'Platinum', 'Diamond', and 'Master' leagues. Dot colours correspond to the player's league.

Effect sizes of the best model remain similar to the original study. Players over 34 years of age tend to be slower by the equivalent of about one league (see Figure 4.1). For actual players, this is a small effect in the sense that it is unlikely that an individual player would notice these changes in their own play, and in the sense that simply practicing more is likely to overwhelm any impact of age. For scientists, this can be seen as a large effect insofar as a league of skill might correspond to hundreds of hours of practice.

As can be seen in Figure 4.1, there is no evidence here that skill in StarCraft 2 reduces the impact of aging. No statistically significant interaction between league and age is found after using a Bonferonni correction to hold the type I family-wise error rate to 0.1. Finally,

adding race to the model results in a better fitting model, but does not change the pattern of results or the estimates of age-related change.

4.1 Developing a Domain-Specific Approximation of Finger Tapping

In this section I will define *RCLatency*, which is based on the median inter-key interval of actions within a screen fixation (see Chapter Two) that contains only right-clicks. There are many screen fixations per game. Therefore, a game's RCLatency is defined as the median of median inter-key intervals described above.

Right-clicks are contextual keys in StarCraft 2 that have a different meaning depending on how they are used. When issued to an empty location on the game map or to a friendly unit, a right-click moves a selected unit. When used on an opponent's unit, a right-click commands selected units to focus their attack on that unit. Finally, selecting a worker and right-clicking on natural resources issues a command to begin mining. In the vast majority of these cases, a second right-click overrides the previous command, making repeated right-clicks redundant (in rare cases the second right-click can be a queued command which will be completed after the previous right-click). 3,326 games have usuable right-click latencies. Repeated strings of right-clicks over very short timespans, which are relatively common (each player has an average of 41.1 screen-fixations containing only right-clicks), do not appear to be cognitively demanding. They consist of clicking on locations to walk, attack, or mine, and they are very rarely the sort of action that requires careful accuracy. I would speculate that the costs of strategizing, decision making, and task switching would occur prior to the first right-click or even prior to the selection of relevant units. Regardless, anyone experienced with StarCraft 2 gameplay would recognize that these clicks can be remarkably fast. The distribution of RCLatency scores (which is noticeably skewed) can be seen in Figure 4.2.

I am treating RCLatency as a prima facie analogue of the finger tapping test, which does show age-related declines (Lezak, 2004) and has diagnostic value in neuropsychological assessment (Lezak, 2004; Schmitt, 2013). It is sometimes assumed that finger-tapping does not track higher level cognitive processes (e.g., Colcombe and Kramer, 2003). One piece of evidence in favour of this view is that people with cognitive deficits, such as in mild dementia, can perform relatively well (Goldman, Baty, Buckles, Sahrmann, & Morris, 1999). Declines in finger tapping resemble changes in the proportion of white matter and, at least on some accounts, peaks around 39 years of age (Batzokis et al., 2011). It makes sense, therefore, to

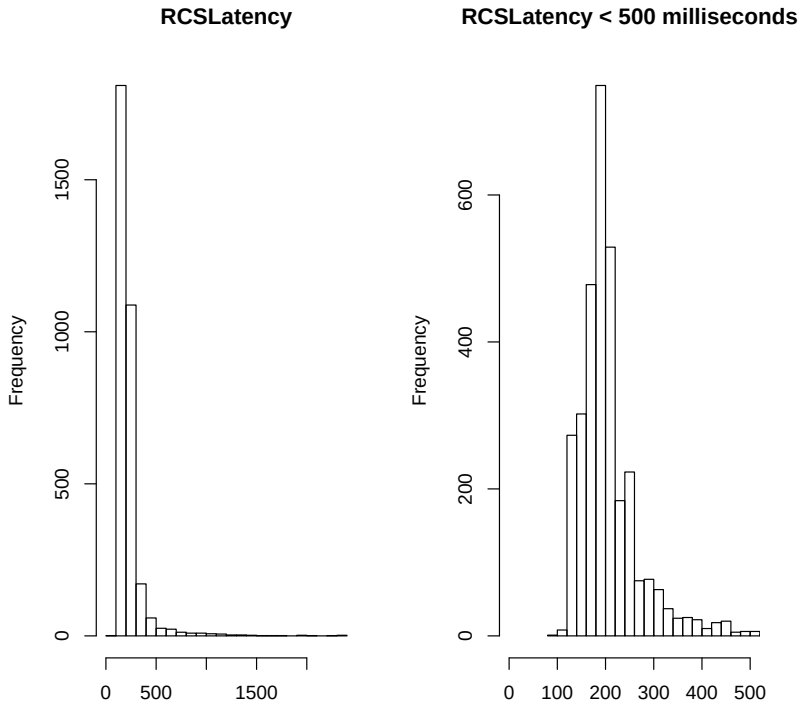


Figure 4.2: Histogram of $RCLatency_i$, which is defined as $median(median(RCLatency_j^{(in\,milliseconds)})$, where i refers to single game and j refers to a single fixation of more than two actions containing only right-clicks.

distinguish the kind of speed present in finger tapping and the speeds that stem from more cognitively complex tasks.[5]

There have also been attempts to construct analogues of the finger tapping test using the inter-key intervals in everyday performance. Austin et al. (2011), for example, measured finger tapping performance using keystroke data from people logging into their computers. In some ways, accessing finger tapping speed in StarCraft 2 is methodologically simpler than in keyboard typing tasks. This is because the atoms in a sequence of keystrokes are not independent, and can be influenced by biomechanical constraints on finger transitions (Loehr & Palmer, 2007), effects associated with transitions between left and right hand

[5]Of course, it is methodologically, and perhaps even philosophically, difficult to judge whether RCLatency measures something *cognitive* in a complex domain. Even the most redundant aspects of StarCraft 2 motor performance, including the unnecessary production of continuous right-clicks, is potentially cognitive when embedded in a task that has relatively clear standards for cognitive assessment. Furthermore, skilled players are *choosing* to perform these redundant actions without instruction, so one might argue that the timings of this performance may still reflect the duration or scope of a cognitive process. All I will argue here is that redundant right-clicks are, at least on the face of things, cognitively simpler than the other behaviours under examination.

(Rumelhart & Norman, 1982), and chunking (Povel & Collard, 1982). RCLatency sidesteps the first two of these methodological concerns by (presumably) focusing on repetitions on the same finger. It deals with the latter concern by focusing on the inter-key interval rather than including the First Action Latency of right-click sequences, as first actions of chunked sequences can be slowed (Povel & Collard, 1982; Rosenbaum, Kenny, & Derr, 1983).

Importantly, RCLatency obviously differs from finger tapping in a number of ways. Note that I will make no attempt to compare RCLatency values to the typical number of taps in the finger tapping test as, even in these tasks, the apparatus can lead to variable results (Lezak, 2004). Furthermore, the finger tapping test usually relies on the index finger (Schmitt, 2013) rather than the middle finger typically used with the the right-click button on a mouse. Also, unlike the finger tapping test, I cannot control for handedness, and this might influence click-speeds for a small number of participants.[6] I can do little to speak to such issues in the present work, and so I simply acknowledge that RCLatency is, at best, a useful analogue of finger tapping.

If RCLatency is a measure of processing speed, then processing speed theory would predict that it would likely explain a moderate or large portion of the age-related differences in mean FAL.[7] However, recall that Salthouse (2000) includes finger tapping as a measure of psychomotor speed that might nevertheless decline alongside other speed measures due to some postulated common cause which impacts aging. Consequently, the present work does not require the assumption that RCLatency is a valid measure of cognitive processing speed in order for it to directly test processing speed theory. If RCLatency is instead a measure of *psychomotor speed* and not *processing speed*, as seems plausible for an analogue of finger tapping, then one might still expect that it will likely account for age-related declines in other variables. The major consequence of this concession, from the purview of processing speed theory, would appear to be that psychomotor speed should not fully explain age-related differences in more complex psychological abilities (Salthouse, 1996), such as mean FAL.[8]

[6]There is some reason to think that left-handed players might exhibit slightly atypical latency values. If they use the mouse with their non-dominant hand, their performance could be impacted. If, on the other hand, they use their mouse with their left hand, they will probably use their index finger, rather than their middle finger, to right-click. In either case, their latencies might be somewhat atypical.

[7]Note, however, that Salthouse (1996) does not insist that every measure of speed will necessarily show age-related decline due to a common cause.

[8]Also recall from Section 4.0.1 that, despite these complications of interpretation, we can still evaluate Salthouse's theory as a tool for explaining performance in a complex task. An earnest effort has been made to use extant theory to explain age-related differences in StarCraft 2 performance and, regardless of the interpretive complications, theories which aid in such an explanation would seem to deserve credit.

Although there is little I can do to settle the issue of whether RCLatency reflects a cognitive process,[9] some exploratory analyses can help me better understand the source of these latencies. The first will consider the dispersion of right-clicks in StarCraft 2 in case that RCLatency reflects not differences in player speed, but in how spatially disperse players make their right-clicks. The second analysis considers whether there are systematic differences between the RCLatencies of fixations with different numbers of right-clicks. Once these possible artifacts influences are removed, I can conduct formal tests of the relations between RCLatency, skill, and aging.

4.1.1 Is RCLatency Biased by the Distance Between Clicks?

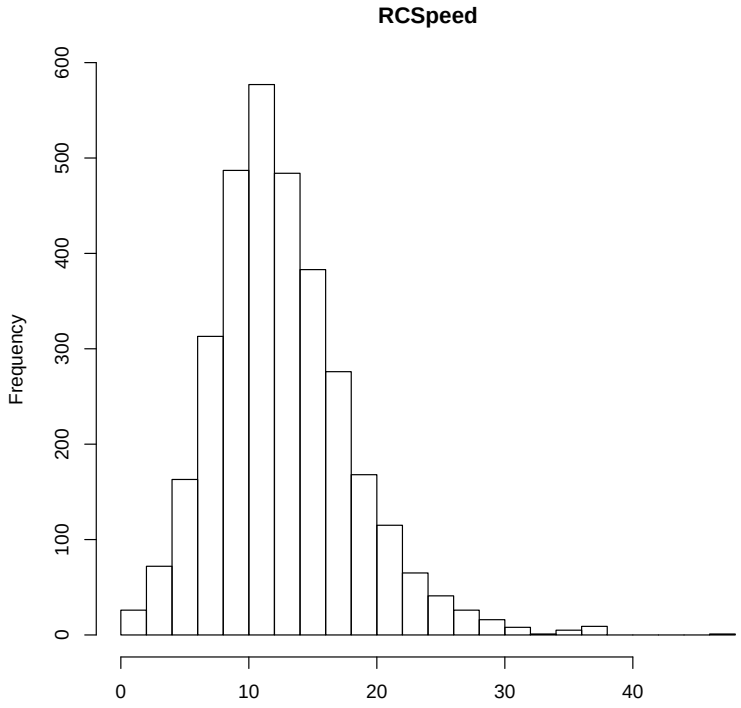


Figure 4.3: Histogram of $RCSpeed_i$, which is defined as $median(\frac{mean(click-distance_j)}{median(RCLatency_j^{(in\ seconds)})})$, where i refers to single game and j refers to a single fixation of more than two actions containing only right-clicks.

I consider the possibility that RCLatency could be biased by the distance between clicks. Fitts' Law, for example, entails that the speed of a right-click is determined in part by the ratio of the distance and width of a target (Fitts, 1954). It would be problematic if players

[9]I take it that settling this issue would likely require a great deal of philosophical or empirical work that is beyond the scope of the present work.

happen to vary the distance between their clicks from fixation to fixation, and even more problematic if certain groups of players preferred less dispersed right-clicks.

I examine how right-click speed varies by Euclidean distance between clicks in individual fixations. I define distance as the mean Euclidean distance of each successive right-click in a fixation. I do not average over all the possible Euclidean distances in a fixation, as I expect right-clicks to get progressively farther from the first right-click (as these are often used to move an army further across the map). Also note that the right-clicks in this analysis only reflect right-clicks with available coordinate information (only 421 right-clicks in the raw dataset lack this information out of 5,167,148 right-clicks in total).

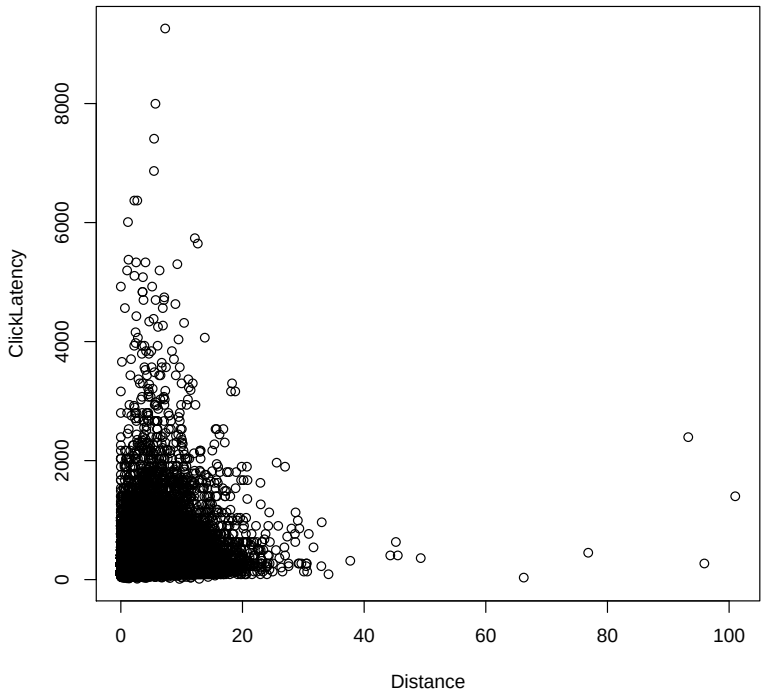


Figure 4.4: RCLatency by Distance. Caution is required here given that each game contributes more than one datapoint (i.e., data are not independent).

Distance values were relatively small (mean=2.81 coordinates, sd=1.4). Thompson et al. (2014) estimate of the view screen size was 40 coordinates by 21.5 coordinates, so distances are relatively small. However, graphical inspection revealed a weak to moderate relationship between RCLatency and Distance (see Figure 4.4). One issue with this approach is that some individuals contribute more fixations, and therefore more data points, to the figure. Comparing each player's median distance to RCLatency reveals a moderate relationship ($r_{Pearson}$= .43, 95% CI [.4,.46]; See Figure 4.5). Furthermore, visual inspection (Figure 4.6) suggested differences in right-click distance by league. Bronze players, in particular,

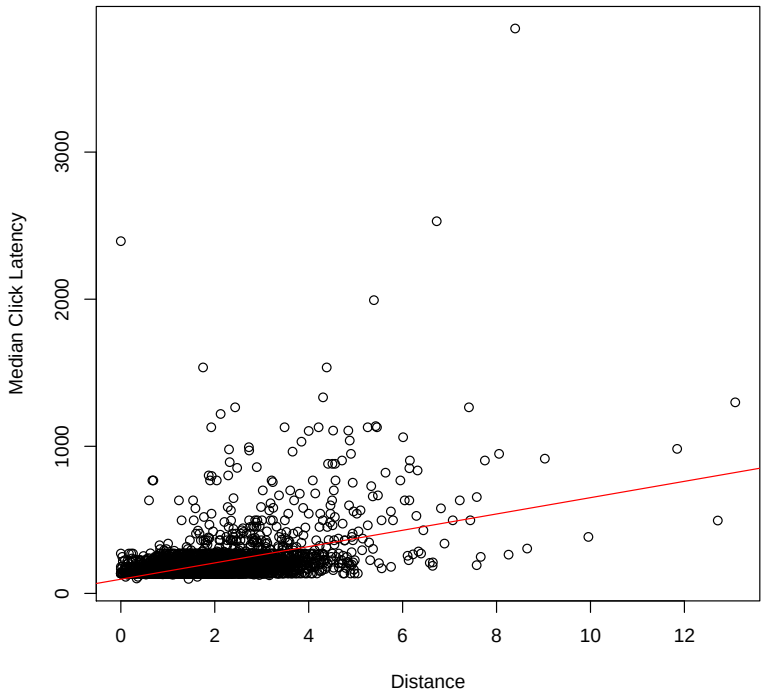


Figure 4.5: Median RCLatency by Median distance between right-clicks.

appear to have larger and more variable right-click distance than other leagues. These results suggest that distance is a potentially serious confound impacting the interpretation of RCLatency. Bronze players might have slower RCLatencies because they are slower generally or because they prefer to click farther apart.

Rather than stick to raw measures of latency, therefore, I also define *RCSpeed* (see Figure 4.3), the median distance in game coordinates between a player's right-clicks divided by the latency of those clicks in seconds (data from 1,427 out of 134,642 had to be dropped from this variable due to having a RCLatency of zero). While this shift from RCLatency to RCSpeed makes a direct comparison with finger tapping speed more difficult, it highlights the fact that, unlike finger tapping, StarCraft 2 right-clicks are usually punctuated by (small) mouse movements. I will retain RCLatency for future analyses as a more direct measure of speed, but will remember the caveat that RCLatency is correlated with the distance between clicks.

4.1.2 Does RCLatency Vary across Different Types of PACs?

RCLatency is an aggregation across fixations with different numbers of right-clicks. This aggregation potentially obscures differences between fixations with a small number of right-clicks and those with many. For example, while fixations containing twenty-five right-clicks almost certainly contains unnecessary actions, a fixation of only two or three right-clicks

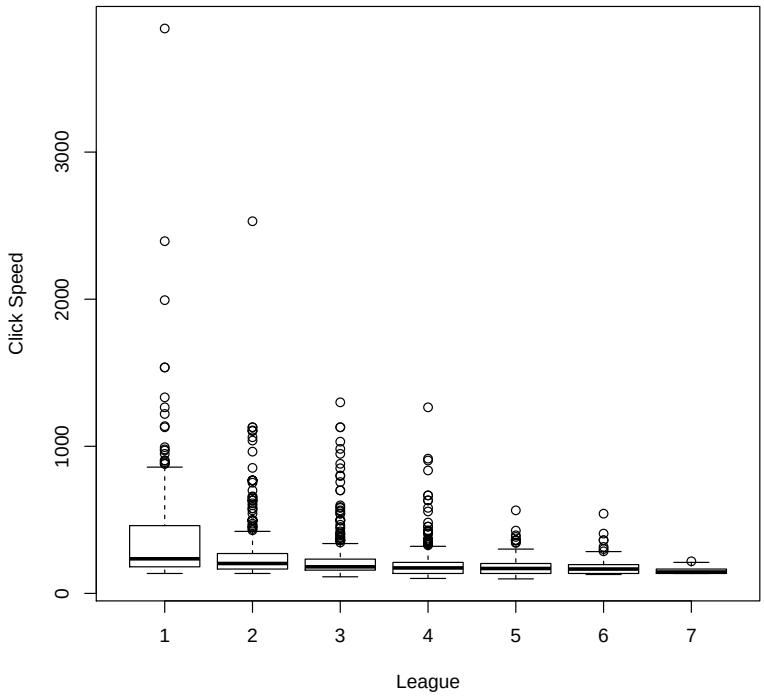


Figure 4.6: Median RCLatency by league.

might reflect a genuine error-correction process. A player might intend, for example, to order their probe to gather minerals but initially misclick and only command their probe to move to a location near minerals. A second right-click, in this case, might reflect the time taken to detect an error and begin its correction. Since aggregating across these two very different tasks would be a mistake, I want to compare RCLatency against the number of actions within a fixation (called FixActionCount).

One difficulty with directly comparing of right-click speed against FixActionCount, and redefining RCLatency over the fastest sort of right-clicks, is that it could lead to a sort of p-hacking, especially if I perform this analysis by player league. It is too easy to simply select whatever definition of RCLatency that exhibits the most impressive or least impressive changes with skill. I instead confine the analysis of FixActionCount to a random subsample (without replacement) of 100 Gold games and 100 Master Games. The results are shown in Figure 4.7. While there may be differences between RCLatency among fixations with three or more actions, fixations with only two right-clicks appear to be unrepresentative. I therefore restrict RCLatency to fixations of at least three right-clicks (Players have, on average, 24.7 such screen fixations per game).

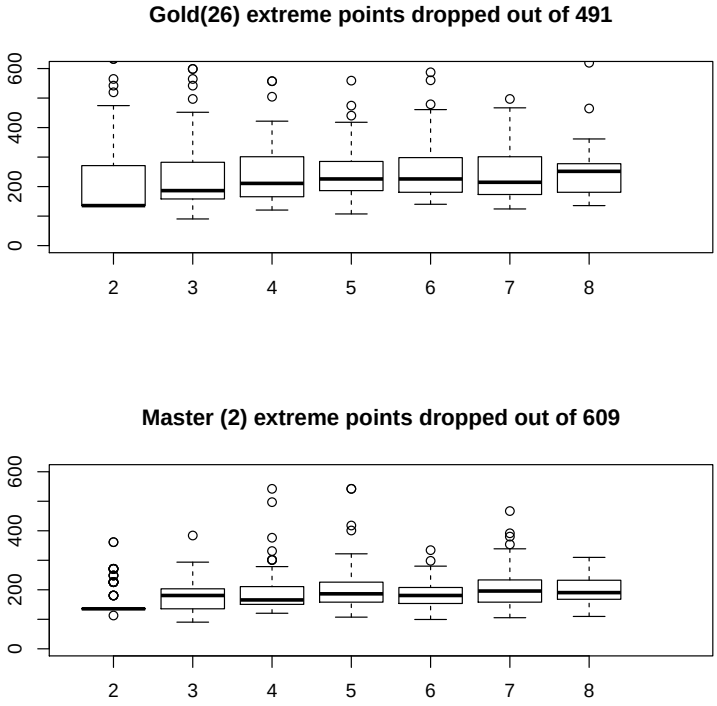


Figure 4.7: RCLatency by PAC actioncount. Each game (n=200) contributes at most one data observation to each level of PAC actioncount. As reported in figure titles, the Y axis is constrained to visualize the bulk of the data and ignore extreme values.

4.1.3 Does RCLatency Differ by Skill Cross-Sectionally?

RCLatency is extremely skewed (see Figure 4.8), and the mean RCLatency is a particularly bad measure of central tendency in Bronze players. Nevertheless, there appear to be differences in the distribution of RCLatencies with skill. A Kruskal-Wallis test over the first six leagues is significant ($\chi^2(5)$=274.07, p<=2.2e-16, η^2=0.0826)[10]. I drop Grandmaster level players here because of low sample size.

I should also note that RCLatency shows a striking floor effect around 100 milliseconds (see Figure 4.8). One should be cautious about inferring that one hundred milliseconds has some special significance here given the issues of resolution discussed in Chapter Two (see table 3.1). Given that latency recordings are tied to the synchronization required by the online portion of StarCraft 2, latency data tends to pile up at zero milliseconds, 90 milliseconds, and 135 milliseconds.

[10] η^2 for all Kruskal-Wallis tests reported here are calculated using guidelines from Tomczak and Tomczak (2014). For comparison, an eta^2 based on a Kruskal Wallis test of mean FAL for the first six leagues is 0.432.

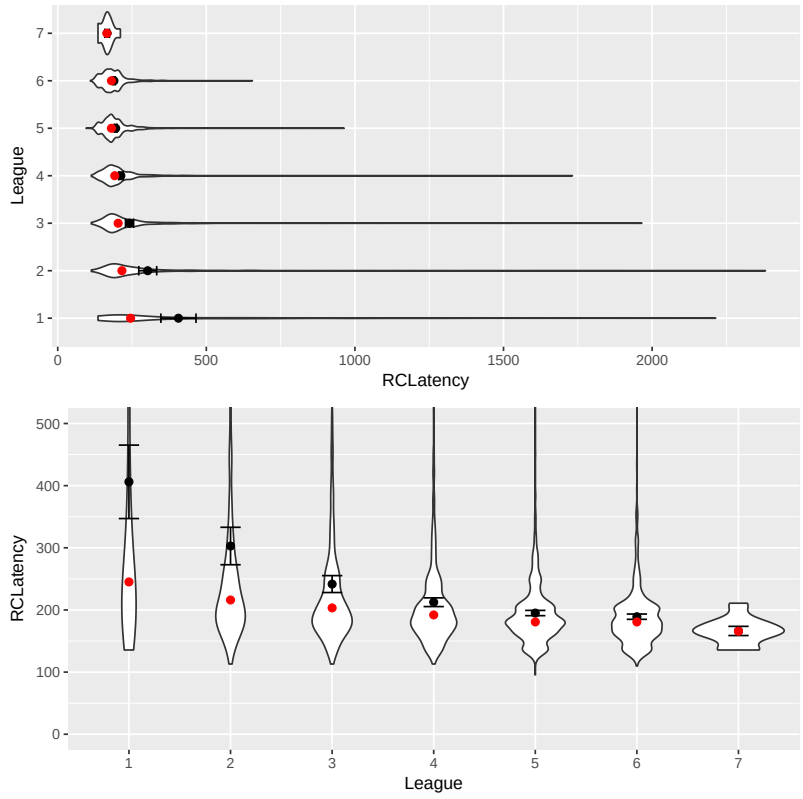


Figure 4.8: Violin plot of RCLatency by league. Black dots represent the mean and whiskers represent two standard errors of the mean. Red dots represent medians.

Figure 4.9 shows the differences in RCSpeed with skill ($\chi^2(5)$=153.29, p<=2.2e-16, η^2=0.045). The distribution is slightly skewed, with the tail of the distribution extending to RCSpeeds of 25 and over, while most leagues have a floor of zero. Interestingly, this floor is elevated to RCSpeeds of two or three coordinates per click latency in Diamond, Master, and Grandmaster leagues. It is worth remembering that, since a game coordinate is extremely small given the 40 by 21.5 coordinate viewscreen on a player's monitor, the RCSpeeds of StarCraft 2 players are not large distances. In short, expert StarCraft 2 players are faster both in terms of the latencies between right-clicks and in terms of how much distance in game coordinates can be covered in the time taken to right-click.

4.2 Does RCLatency Account for Declines in Mean FAL?

4.2.1 Analytic Strategy

I begin by examining whether RCLatency exhibits age-related differences. I will use likelihood ratio tests to evaluate the following models of increasing complexity.

$$Model_1^{RCLatency} : RCLatency \sim (league * age)$$

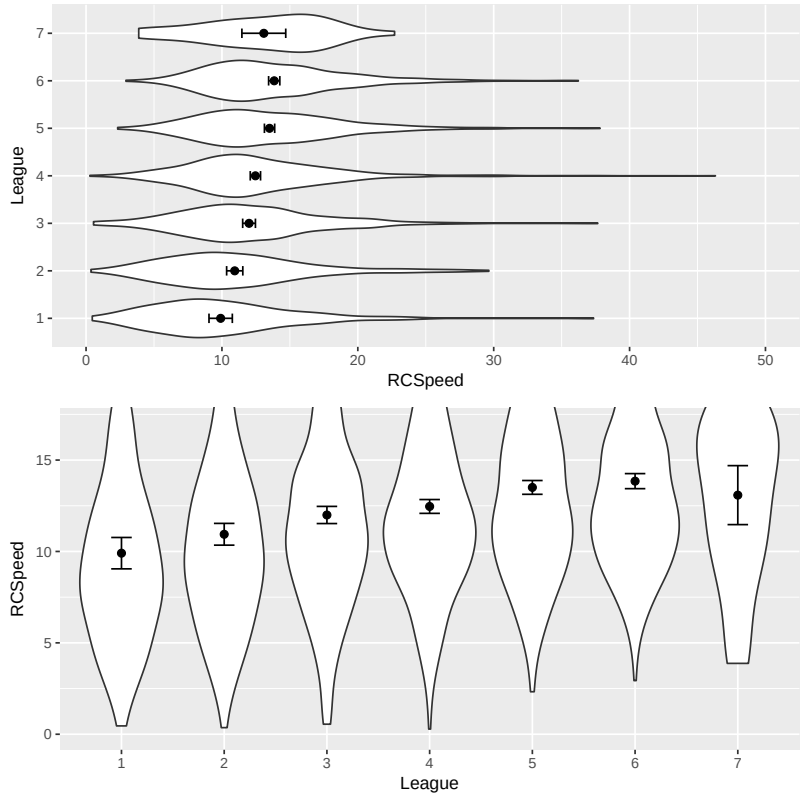


Figure 4.9: Violin plot of RCSpeed by league. Black dots represent the mean and whiskers represent two standard errors of the mean.

$$Model_{seg}^{RCLatency} : RCLatency \sim (league * age) + age_{segmented}$$

As in the previous analysis, I evaluate these models using a likelihood ratio test.

Further analysis is needed to establish whether RCLatency can explain previously observed relations with mean FAL. I follow Salthouse's (1996) formalization of this question in terms of the proportion of age-related variance in mean FAL that is shared with RCSpeed. I am interested in whether RCLatency (or RCSpeed) can explain the variation represented in $A \cup B$ from Figure 4.10. Note that I am *not* intending to explain all of the variability in mean FAL. I acknowledge that a lot of the variation in mean FAL[11] will not be explained in the present work. The primary concern here is the extent to which the variability associated with region B (which represents variability that is shared between age, mean FAL, and RCLatency) takes up a large proportion of the variability associated with $A \cup B$ (which represents shared variability between age and mean FAL).

One of Salthouse's (1996) statistical approaches to this problem is to compute squared Pearson correlations to represent the A∪B region and squared semi-partial correlations to

[11]This variability might be due to a wide variety of factors such as differences in motivation or decision making habits.

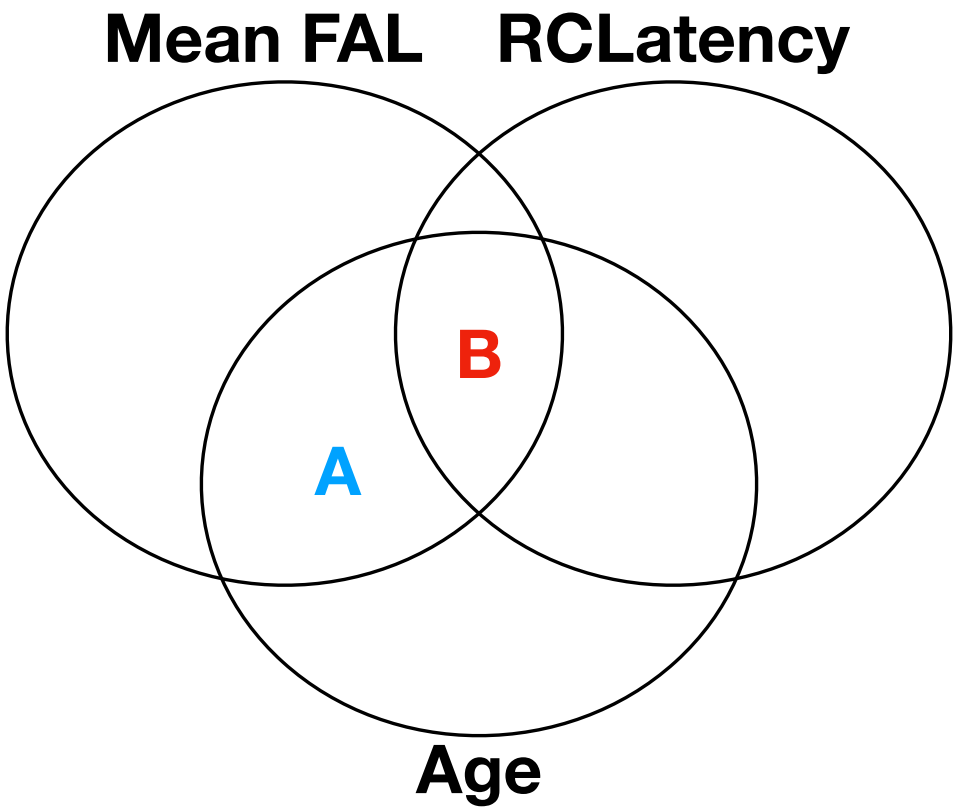


Figure 4.10: Venn diagram representing variance and shared variance between mean FAL, RCLatency, and age. The region of $A \cup B$ corresponds to the variability we would like to explain. I take processing speed theory to predict that region B will occupy a large proportion of $A \cup B$.

represent the A region. This allows one to calculate the proportion of age-related variance in mean FAL that is shared with RCSpeed (which represents the A region).

$$B = \frac{r^2_{pearson} - r^2_{semi-partial}}{r^2_{pearson}}$$

I expect this proportion to be large if RCSpeed accounts for the age-related declines in mean FALs.

Given that the relationship between aging and mean FAL appears linear after the age of 24, I consequently restrict the analysis to ages 24 and older. After this restriction, relationships between age, mean FAL, and RCSpeed are approximately linear without clear breakpoints. I therefore proceed to calculate semi-partial correlations based on the residuals of a simple linear regression and Pearson product moment correlations.

4.2.2 Results

Aging in RCLatency

To probe RCLatency for aging effects, I began by comparing the base model

$$Model^{linear} : log(RCLatency) \sim age * league$$

against its segmented counterpart. Issues with heteroscedasticity required log-transforming the dependent variable. However, the likelihood ratio test revealed no evidence that the segmented model was an improvement over the base model ($Model^{linear}$ =13, $Model^{segmented}$ =15, $\chi^2(2) = 1.59, p = .45$).

Given that the relationship between Log RCLatency and age appears approximately linear (and a linear regression of log RCLatency on age has sensible looking residuals), I computed a Pearson correlation to describe the strength of the relationship. The relationship appears to exist, though it is very weak ($r = .11, 95\%CI[0.08, 0.14], t(3000) = 6, p = 4 * 10^{-10}$).

The strength of the relationship between age and log mean FAL after 24 was $r_{pearson}$= .35 (95% CI [0.29 0.41], n=888), and the partial correlation between age and log mean FAL, controlling for the effect of RCLatency on age, was $r_{partial}$=0.31 (95% CI [0.25 0.37]).[12] This suggests that the proportion of age-related variance in mean FAL that is shared with RCLatency ($\frac{r^2_{pearson} - r^2_{semi-partial}}{r^2_{pearson}}$) is about 19%. In order to test that this percentage of shared variance is non-zero, I used the Pearson and Filon (1898) test against equality of correlations with dependent and overlapping groups, implemented in the R package *Cocor* (Birk Diedenhofen, 2016). The null hypothesis is rejected, and I accept that the full correlations (between age and median FAL) and partial pearson correlations (between mean FAL and the residuals of $age \sim RCLatency$) are distinct (z=10.94, p<= 0.0001, n=888).

The picture does not change substantially if I instead consider the proportion of age-related variance in mean FAL that is shared with RCSpeed (13.5%, z=10.74, p<= 0.0001, n=888), suggesting that these results are not due to differences in how older and younger individuals space their clicks.

4.2.3 Rerunning Analysis Using Median FALs

In order to ensure these results were not due to the selection of mean FALs as opposed to median FALs, I reran the analysis using median FALs. The medians simplified the modelling process somewhat as log-transformation of the dependent variable was not required in order to address issues of homoscedasticity. The resulting model had a a slightly older onset of aging, at 27 years (with a confidence interval of 24-30 based on two standard errors). This segmented model outperformed the nested model $medianFAL \sim age * league$ model

[12]The Pearson correlation between log mean FAL and RCLatency was .4

($Model^{linear}$ =13, $Model^{segmented}$ =15, $\chi^2(2)$=11.14, p=0.004). For a fuller descrption of these results see Appendix ii.4. These differences are interesting and, given the sensitivity of the mean to extreme values, presumably due to differences in how the distribution of FAL scores differ across age which are independent of central-tendency. While it is beyond my scope here, future work should consider analyses that which would explain why the analysis of means and medians give different results.[13]

I also asked whether RCLatency could explain effects of age on median FAL, replicating my previous analysis except with data from 268 individuals who were 27 or older (also see Appendix ii.4 for additional details). The results were a $r_{pearson}$ = .36 (95% CI [.25 .46]), a $r_{partial}$ = 0.29 (95% CI [.18 .4]). There was also evidence that the two correlations were different (z=6.66, p=0.0001). Consequently, about 36% of the age-related variance explained by median FAL appeared to be shared with RCLatency (i.e., $B = \frac{r^2_{pearson} - r^2_{semi-partial}}{r^2_{pearson}}$=.36). In summary, the capacity of RCLatency to explain aging effects seems to be somewhat larger if the analysis is restricted to median FALs, but the overall pattern of results changes little. RCLatency does appear to explain some small proportion of the age-related differences in First Action Latencies.

4.3 Discussion

My primary objective was to further probe the effects of age-related differences in mean FALs found by Thompson et al. (2014). I confirmed those results by rerunning the data using a more appropriate statistical technique that was specifically designed for the creation of segmented models, and proposed a simpler measure of response speed, RCLatency, which tracks the speeds at which players right-click with their mouse. The findings suggest that, while RCLatency does explain some of the age-related declines in mean FAL, the proportion of age-related variance in mean FAL that is shared with RCLatency is only about 19%.

The results are not due to age-related differences in the distance between the right-clicks of players, as the results change are roughly the same if I use median RCSpeed instead of RCLatency. Indeed, the shared-age related variance in mean FALs that is explained by RCSpeed is only 13.5%. I also ruled out potentially unrepresentative fixations using a pilot subsample of Gold and Master players.

It is important to remember that the present work constitutes an application of processing speed theory to an entirely new domain. Unlike studies of typing skill, which are observations upon which processing speed theory offers a hindsight explanation (Bosman, 1993; Salthouse, 1984), there has been no opportunity to create ad hoc adjustments of the

[13]Thanks goes to Jack Davis for proposing one such an analysis. One could rerun the aforementioned analyses not only with median FALs but with a variety of percentiles (e.g., the 60th, 80th, and 90th percentile of FAL scores), to see whether age-related differences in FAL might begin with a fattening of tails rather than a difference in central-tendency.

theory in order to account for StarCraft 2 data. Nevertheless, I found, as processing speed theory would predict, that statistically significant proportions of age-related variance in one response time measure will be shared with speed measures that appear to tap into entirely different cognitive capacities. This testifies to the robustness of processing speed theory predictions. However, I find no evidence that a general factor might explain a *large* proportion of the age-related differences in speed. Recall that this is consistent with Salthouse (1996) insofar as RCLatency is thought of as a measure of raw psychomotor speed. Salthouse (1996) argued that the age-related differences in speed could not be explained by psychomotor speed alone.

Finally, I reran the entire analysis using median FAL instead of mean FAL, obtaining a similar pattern of results. The choice in dependent variable does impact estimates of the proportion of age-related variance that is shared with RCLatency (Location B from Figure 4.10 corresponds to roughly 19% using mean FAL and 36% using median FAL), but it is important to remember that these analyses are predicated on segmented models of when age related declines begin. The onset of aging is slightly later for median FALs (27 years as opposed to 24), and this impacts the sample sizes contributing to the estimate of section B in Figure 4.10. Regardless of how the analysis is conducted, the proportion of shared age-related variance is relatively small but non-zero.

One key limitation of this study still requires attention. I will interpret these results in light of debates about the use of cross-sectional methods in studies of age-related differences.

4.3.1 Issues with the Use of Cross-Sectional Data

There is an ongoing debate about whether cross-sectional findings can be used to make claims about cognitive declines, particularly when cross-sectional studies and longitudinal studies often differ in their estimates of when aging begins (Salthouse, 2009). Schaie (1965, 2009) has argued that cross-sectional differences cannot be used to infer longitudinal change in these cases, often by pointing to issues of cohort effects. Salthouse (2009) has responded that such accounts need to do more to specify exactly what features of cohorts could account for observed differences. 'Some variables, such as years of education, are easily assessed, but the prediction that cross-sectional age trends would be eliminated after adjusting for these other variables cannot be adequately tested until all of the cohort-relevant variables are identified and measured' (p. 7). The implicit premise here appears to be that the burden of proof lies with critics of cross-sectional investigations of aging to specify measurable nuisance variables which could account for aging effects. A second approach to defending cross-sectional studies of aging is to point out that longitudinal studies have their own limitations, including cohort effects (which are potentially confounded with age where one is unable to sample from multiple cohorts) and retest effects (Salthouse, 2009).

I do not wade into this debate here. Instead it is taken for granted that cross-sectional differences in StarCraft 2 performance could be due to any combination of cohort effects and genuine age-related differences.

4.3.2 Closing Remarks

The primary contribution of this chapter is that, while the timings of age-related difference in StarCraft 2 map relatively well to the ages associated with peak performance in male 10,000 meter runners, who peak at about 23 years of age (Berthelot et al., 2012), the age-related declines in mean FAL appear to have more complex underpinnings than a simple decline in the ability to tap fast: an ability tracked by finger tapping test (Lezak, 2004). If a clicking speed decline did underlie these results, I would expect RCLatency, a measure of performance speed in redundant clicks that place fewer cognitive demands on participants than the typical FAL, to show the clearest declines with age. On the contrary, RCLatency appears to show only weak age-related decline, and it explains little of the age-related differences observed in mean FAL. These results are more consistent with age-related slowing in planning, perceptual speed, or complex motor sequence preparation.

However, there are independent reasons for thinking that differences in mean FAL are not due to declines in complex motor sequence preparation. If mean FALs were heavily influenced by sequence preparation, then I would expect slower First Action Latencies in screen fixations which contained many actions. However, Thompson, McColeman, et al. (2017) found only weak spearman correlations between the number of actions in a fixation and First Action Latency, and this finding required the ad hoc exclusion of fixations containing only a single action. My best guess, therefore, is that differences in mean FALs should be thought of as being likely due to differences in perceptual or planning speeds.

Chapter 5

The Relevance of Big Data to the Psychology of Complex Skill Learning

There is some reason to think that the relationship between Big Data and psychology will be somewhat different from the relevance of Big Data to other domains. Unlike disciplines such as business and commerce, which are sometimes forced by their subject matter to deal with increasingly troublesome datasets, psychology is not compelled to embrace these new data sources. Psychology has the laboratory, and there are enough questions to be asked about performance on basic laboratory tasks for psychology to continue using traditional methods indefinitely. Furthermore, Big Data brings a number of methodological and statistical challenges. Indeed, Laney (2001) defines *Big Data* in terms of its unwieldiness. On this definition, Big Data refers to data which are either too large for traditional processing (volume), is so heterogeneous with respect to data types that analysis becomes extremely awkward (variety), or is updated so frequently as to be problematic (velocity).

Understanding the relevance of Big Data to psychology, therefore, requires a discussion of whether the scientific advancement promised by emerging data sources could offset the methodological challenges it entails. This means that the ensuing discussion will actually be about the previously untapped behaviours *represented* in Big Data, not about the volume, variety, or velocity of data. Of course, since many of these new data sources can be sampled from, rather than analyzed in their entirety, many researchers will still be able to exploit these new data sources without necessarily using Big Data (Goldstone & Lupyan, 2016). For the remainder of the chapter, therefore, *my goal will be to explain how psychology might exploit emerging data sources for scientific advancement.* What I argue will ultimately apply to Big Data as well, since fully exploiting naturalistic telemetry will likely mean coming into contact with Big Data.

Chapters Three and Four have prepared me, at least somewhat, to address the question of Big Data's potential value for psychology. The longitudinal dataset utilized in Chapter

Three was arguably Big Data as it was large enough to be methodologically challenging. The typical methods that were used for processing data on a single local machine (and storing them on a small number of MySQL tables) was acceptable for the 32 million row cross-sectional dataset. The method needed to be rethought for the longitudinal dataset of about one billion rows, which required parallel processing and the distribution of raw data over a large number of MySQL tables. The cross-sectional dataset used in Chapter Four would probably not constitute Big Data for many researchers, but it also deserves consideration as it nevertheless exploits an emerging data source that has yet to be fully utilized in psychology.

After reviewing the contributions of Chapter Three and Four, I address an unfortunate complication in any discussion about the relationship between Big Data and psychology. Some have argued that Big Data allows researchers to sidestep theory, and even *replace theory* (C. Anderson, 2008). At the heart of such arguments is the view that Big Data will radically transform our fundamental approach to research, with data-driven approaches essentially replacing hypothesis-driven science. Understanding the relevance of Big Data to psychology, therefore, also requires addressing these beliefs.

The chapter will continue by tackling common responses to Anderson's (2008) argument (Section 5.1). I take these responses to be devastating to the notion that Big Data could replace theory, and so I will not attempt to levy additional critiques here. Instead, Section 5.2 turns to what I see as the kernel of truth in this argument. I will argue that the variety of new data sources, while perhaps not a threat to the need for *theory*, is nevertheless a threat to the particular theories we presently endorse. That is, while Big Data cannot do away with the general importance of theory, the variety of new data sources, many of which capture complex performance outside of the lab, nevertheless threaten to overturn theories which make overzealous generalizations. In 5.2 I argue that the capacity to identify problematic generalizations might profoundly change research practices. The increased capacity to test theory generalizability will afford researchers more opportunity to back up their generalizations and, insofar as this opportunity is ignored, overzealous generalizations are likely to constitute bullshit (Frankfurt, 2009).

Having reviewed the contributions of Chapter Three and Four, and having articulated the utility of Big Data in examining theory generalizability, the rest of the chapter turns to two specific research strategies that utilize emerging data sources to examine theory generalizability. Two obvious possibilities present themselves. I will call this the *Model-Organism* approach and the *Free-For-All (FFA)* approach.

5.0.1 A Review of the Contribution of Chapters Three and Four

The contribution of Chapter Three was the identification and addressing of a new source of complexity in the complex skill StarCraft 2. While most learning in the laboratory can be sensibly studied by operationally defining 'experience' in terms of a single variable (I

call this the presumptive definition), such an approach is potentially misleading in the case of StarCraft 2. On the uni-dimensional approach one would probably, like myself, begin by defining experience in terms of 1v1 competitive experience, as this is the context that professional players inhabit. However, my results suggest that this presumptive definition is misleading in two ways.

First, the presumptive definition is misleading because it disregards race-specific learning. In most contexts, models were improved[1] by adding additional information about race experience. In the initial analyses, this additional information took the form of variables describing the number of Protoss and Terran games played. In the analysis of only dominant-race play, this additional information took the form of a variable describing how many off-race games are played.[2]

Secondly, the presumptive definition is misleading because it ignores other sorts of experience that do not fall within 1v1 gameplay. In all analyses considered here, a variable describing 2v2 experience seemed to improve model performance (according to the likelihood ratio test). Furthermore, examination of the coefficients for our final models (see Figure 3.16 and Appendix ii.2 for details) suggests that the impact of 2v2 experience, while perhaps difficult to explain, is not the same as 1v1 experience.

Some of my findings are not surprising given research on skill transfer. Indeed, if experience in shooting basketballs from the foul line does not transfer to skill in the same behaviour over slightly larger distances (Keetch et al., 2008), then it should not be a surprise that skill in 2v2 StarCraft would not have the same effects as 1v1 play, as the two game modes are quite different in terms of how attacks are timed and in terms of what strategies are effective. These findings are also not irreconcilable with theory, even the classic picture of skill as a hierarchy of habits (Bryan & Harter, 1897, 1899), as it is plausible that different game modes require slightly different habits or skills. We would predict that practice which only trains a subset of the required habits will be less effective.

In addition to showing that a uni-variate approach to defining experience is potentially misleading, Chapter Three also estimated the extent to which skills transfer. For example, if we hold all other variables constant and restrict ourselves to a player's dominant-race, a game of off-race experience appears to *slow* speed by about 23% of the value that one game of dominant-race experience would improve speed. In contrast, one game 2v2 experience appeared to be suspiciously beneficial, conferring 165% of the speed up predicted by a

[1]Recall I judge improvement here based on likelihood ratio tests

[2]The only analysis where off-race experience did not appear to lead to improved model performance was where the dataset was reduced to the 33 players who had more than 30 2v2 games. I do not take this too seriously as the goal of this followup analysis was to explain the seemingly large effects of 2v2 experience. Nevertheless, even if we take this result seriously, it does not cast doubt on the larger point that uni-variate definitions of experience are misleading because, even in this analysis both 1v1 experience and 2v2 experience appeared to play an important role.

single game of 1v1 experience.[3] Both of these latter results are somewhat surprising, and presumably could not have been found without a direct inquiry into many hours of StarCraft 2 play.

In short, Chapter Three brings psychology one step closer to Newell's (1973) goal of a complete explanation of performance in one complex domain. In doing so it also offers insight into how researchers might begin inquiries into other complex skills of more direct societal interest. The most prominent of these insights is that researchers studying a complex skill should at least consider a multi-dimensional operationalization of experience and, ideally, to allow data to help inform these decisions. Otherwise, they may wind up relying on common-sense definitions of experience, such as 1v1 experience,

Chapter Four had two contributions, one empirical and one theoretical, and both of these contributions stem from extending the work of Thompson et al. (2014). Previously, cross-sectional studies using simple laboratory tasks had suggested that an effect of aging on performance could be seen at relatively young ages (Salthouse, 2009; Schroeder & Salthouse, 2004; Tsang & Shaner, 1998). Thompson et al. (2014) found analogous age-related differences beginning at 24 in the response speeds of StarCraft 2 players. This suggested that declines found in the laboratory may manifest themselves in much more complex behaviours. These results were surprising as, in complex tasks, experts are able to use the regularities of their environment in order to circumvent many such low-level changes. Older typists, for example, appear to be able to avoid age-related declines in typing speed by reading ahead (Bosman, 1993). Chapter Four's empirical contribution lay in answering whether the age-related differences in response time could be explained by a much simpler behaviour, redundant Right-Click Latencies.

Our results suggest that variability in RCLatency can only account for about 19% of the shared age-related variance in age and mean FAL. This suggests that we probably should not think about these age-related differences in mean FAL as being due to the ability to click fast.

A further contribution involves the evaluation of Salthouse's *processing speed theory*. The observed pattern of correlations are consistent with the existence of a common cause which declines with age and impacts both RCLatency and mean FAL. However the small size of this effect suggests that, even if a common cause could be identified and directly measured, the age-related variance in mean FAL would remain largely unexplained. This is consistent with processing speed theory if we are willing to accept that RCLatency is a measure of psychomotor speed, which Salthouse (1996) does not think can fully explain

[3]I can not currently explain why 2v2 games might be so valuable to StarCraft 2 players. The most likely possibility is that players who play these game modes benefit a lot, perhaps because they are weaker players with more to learn. The followup analyses considered here only examined player dedication and the number of 2v2 games played. Future work should consider whether these findings can be explained by independent estimates of player skill, such as MMR.

age-related differences in other speed measures. Furthermore, the theory has been of at least some practical utility in explaining results from studies of complex performance in a novel domain (Thompson et al., 2014).

5.1 Objections and Responses to Concerns about Big Data

One exaggeration[4] of Big Data's importance that has received a lot of attention comes from *Wired* magazine (C. Anderson, 2008):

> 'In short, the more we learn about biology, the further we find ourselves from a model that can explain it.
>
> There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.' (¶12-13).

Anderson is careful to distinguish correlation from causation, but denies such understanding is necessary for practical purposes. Insofar as my goal is to maximize quarterly profit, for example, one might do better predicting profit with an opaque deep learning algorithm than by appealing to my best theory. If Anderson were merely claiming that, for practical purposes, researchers are sometimes content with identifying correlations in the absence of understanding causation, then he is in good company and certainly not making some fundamental attack on theory. However, Anderson seems to go farther and instead concludes that '[... Science] can advance even without coherent models, unified theories, or really any mechanistic explanation at all' (C. Anderson, 2008, ¶19). This claim suggests that Big Data can allow for scientific advancement even outside of those restricted cases where predictive validity is sufficient for research purposes.[5]

Direct rebuttals of Anderson and similarly worded claims from the media are common place (Bollier, 2010; Boyd & Crawford, 2012; Mazzocchi, 2015). In some cases responses include obvious points which, in other contexts, might come across as pedantic.

[4]We mention less contentious interpretations of these claims below, but attention to their contentious interpretation is of more value to the present argument.

[5]Interestingly, these comments appear to have touched a nerve even among empirical researchers in cognitive science as one can even find attempts to highlight foundational importance of theory in otherwise empirical work.

> Rather than *supplanting* laboratory investigations, research using [Big Data or naturally occurring datasets] can supplement traditional approaches by serving as a proving ground for theories developed in rigorously controlled experiments (emphasis added, Paxton and Griffiths, 2017, p. 1).

> Extrapolating from correlations can yield specious results even if large data sets are used. The classic example may be "My TiVO Thinks I'm Gay." The Wall Street Journal once described a TiVO customer who gradually came to realize that his TiVO recommendation system thought he was gay because it kept recommending gay-themed films (Bollier, 2010, p. 5).

The critique here is that Anderson underestimates the practical utility of theory. While some research questions might be answered solely with correlations, there are many practical cases where a myopic focus on correlation could lead one astray.

Perhaps more interestingly, Anderson's claims are seen as an opportunity to delve into foundational issues in the philosophy of science.

> At present, whilst it is clear that Big Data is a disruptive innovation, presenting the possibility of a new approach to science, the form of this approach is not set, with two potential paths proposed that have divergent epistemologies - empiricism, wherein the data can speak for themselves free of theory, and data-driven science that radically modifies the existing scientific method by blending aspects of abduction, induction and deduction (Kitchin, 2014, p. 10).

I read Anderson as electing a theory-free sort of empiricism[6] when he claims that '[we] can analyze the data without hypotheses about what it might show' (C. Anderson, 2008, ¶13).[7] However, insofar as making use of correlations requires knowledge about what measures mean, Big Data cannot be free from theory. For realists about measurement, interpretation of measures logically requires some consideration of the mind-independent entities

> The rapidly increasing availability of high-quality data logging, storage, querying, visualization, and analysis tools means that the bottleneck for progress in the social sciences *increasingly on the theoretical side*. Cognitive science has always been strong on theory, and its practitioners have unique theoretical vantage points from which to explore real-world data sets. If theory is indeed the bottleneck, then cognitive science becomes a perspective of increasing societal relevance because of its ability to widen the constriction (Goldstone & Lupyan, 2016, p. 565).

This salute to theory is surprising in the context of empirical papers, where foundational issues in the philosophy of science are rarely mentioned. Researchers exploiting novel data sources seem to find themselves distancing themselves from claims such as Anderson's (2008).

[6]Theory-free empiricism seems extreme, even for practical purposes. If we intend to use correlations for industry applications, theory is presumably required to judge whether findings are actionable. If office productivity is higher in summer months, data scientists are not obligated to increase the office temperature year-round. I avoid these responses here in order to avoid debates about whether Actionability could be estimated within a theory-neutral statistical model.

[7]One possible reading is that Anderson is just advocating data-driven research and induction. If so, then his claims are not new. Indeed, Mazzocchi (2015), for example, points out that the excision of theory-driven research resembles Bacon's proposal of a scientific method relying heavily on induction over deduction (Standford Encyclopedia of Philosophy, 2003)

which causally produced their datasets (*Standford Encyclopedia of Philosophy: Measurement in Science*, 2015, §1). The measurements of non-realists, on the other hand, will at least need to be accompanied by a theory of what each datum indicates about the world (*Standford Encyclopedia of Philosophy: Measurement in Science*, 2015, §8.2). To the extent that all measurements are theory-laden, the proposal of a theory-free empiricism is a non-starter.

Unfortunately, the responses to Anderson are too numerous and cutting for me to make a novel contribution by levying yet another critique. Therefore I will come to his defence instead. I will not argue that previous critiques have been uncharitable or misguided. Instead, I will argue that there is a kernel of truth to the notion that Big Data is a major threat to psychological theory. I distinguish between the need for *theory* generally, which encompasses the need for at least some interpretation of what numbers mean, and the particular theories we happen to endorse today. Big Data is no threat to the need for *theory* generally, but has the potential to threaten particular theories which are so general in scope as to apply to a wide array of complex behaviours. Big Data may also inspire some healthy modesty about our current understanding of the world and in the capacity of our laboratory studies to generalize.

Importantly, I will not argue that Big Data will dramatically reshape psychological theory in the short term, though it is possible that this is the case. Instead, I first explain how the emerging variety of data sources, many of which will qualify as Big Data, offer new opportunities for testing the generality of theory. Later, I will argue that these new opportunities have the potential to shift community standards surrounding the pronouncements of theory generalizability.

5.2 Data Source Variety and Generalizability of Specific Theories

The threat of Big Data to theory does not come from the size of datasets. As Boyd and Crawford (2012) note, census data is often larger than the datasets employed by Twitter researchers. Instead, I will argue that emerging data sources can pose serious questions for theories whose generalizability was previously unquestionable.

One of the most important emerging sources are Naturally Occurring Datasets (Goldstone & Lupyan, 2016). Our lives are increasingly computer mediated, and so complex human behaviour is increasingly logged in freely available databases.[8] Consequently, theories which have previously been free to purport some degree of generality are increasingly having this generality tested against emerging data sources. Today, a complete theory about

[8] Insofar as these data are recorded on company or participant computers rather than by the researcher, naturalistic telemetry will usually constitute a Naturally Occurring Dataset. I therefore use the terms interchangeably here.

relationships and social networks must extend to the hundreds of million of tweets per day (*Internet Live Stats: Twitter Usage Statistics*, n.d.). Those theories which make claims about how cognitive-motor performance changes with time are increasingly evaluated in light of video games. For example, Thompson, McColeman, et al. (2017) examined the generalizability of theories of motor execution and Thompson et al. (2014) examined whether cross-sectional declines in basic laboratory speeds (Salthouse, 2009) generalized to a much more complex domain.[9]

It is crucial to note that, although theories which purport to be relevant to complex human behaviour are increasingly coming into contact with new data sources, tests of generalizability are not easy or necessarily definitive. Twitter research has a number of methodological challenges, such as the difficulty in getting a random sample of tweets when Twitters API does not give access to the full database, and when some tweets are censored (Boyd & Crawford, 2012). It also may turn out that traditional theories cannot generalize to Twitter because interactions on Twitter operate under fundamentally different constraints (Boyd & Crawford, 2012). Consequently, just because a theory does not appear to generalize to a novel domain, the failure may be due to domain specific differences or methodological complications.

Similar challenges were encountered when I investigated cognitive-motor speeds following a major shift of attention in video game players (Thompson, McColeman, et al., 2017). Actions following attentional shifts were delayed, which would be expected if the actions following attentional shifts were motor-chunks (Povel & Collard, 1982; Rosenbaum et al., 1983; Yamaguchi & Logan, 2014). However, Thompson, McColeman, et al. (2017) found only weak evidence for the view that executing longer sequences of actions exacerbate these delays. This was incompatible with predictions that that there should be a latency cost for loading additional actions into a a motor buffer prior to execution (Henry & Rogers, 1960; Sternberg, Monsell, Knoll, & Wright, 1978; Yamaguchi, Logan, & Li, 2013)). This effect was only found after the exclusion of attentional shifts that were followed by only one action and, even then, correlations between response time and sequence size were weak. Although more research is of course required to call this a strict failure of theory to generalize, it seems clear that data sources from video game players have extended our understanding. If nothing else, the burden is now on supporters of a general motor buffer theory to provide guidance about how to explain this aspect of complex behaviour.

Of course, I have only provided a few examples of how naturalistic telemetry can be a powerful tool for evaluating the capacity of theories to generalize to behaviour outside of the lab. I don't intend to specify the extent to which naturalistic telemetry will revolutionize theory evaluation, but I do want to explain how a qualitative shift in the academic landscape

[9]For further discussion on the value of Big Data and naturally occurring datasets in testing laboratory theory, see Paxton and Griffiths (2017).

would be possible. I am imagining a state of affairs where naturalistic telemetry allows us to think more critically about the purported generality of theories. That is, we will be able to distinguish between theories which provide guidance to those studying naturalistic telemetry and those whose pronouncements of generality should not be taken seriously.

5.3 Naturalistic Telemetry as a Frankfurtian Bullshit Detector

I argue here that naturalistic telemetry, and naturally occurring datasets in particular, could serve as a Frankfurtian bullshit detector. According to Frankfurt (2009), bullshit can be distinguished from lying because the liar is necessarily concerned with the truth (insofar as their aim is to deceive someone about it). A core feature of bullshit, although it might be designed to give a specific appearance (e.g., the appearance of legitimacy, wisdom, strength, etc.), the wielder need not have any concern for the truth of what they say. What makes bullshit undesirable is not that bullshit is necessarily false, but that bullshit implies a disregard for the expectation of due care which appears to be implicit in polite conversation.

> What bullshit essentially misrepresents is neither the state of affairs to which it refers nor the beliefs of the speaker concerning that state of affairs. Those are what lies misrepresent, by virtue of being false. Since bullshit need not be false, it differs from lies in its misrepresentational intent. The bullshitter may not deceive us, or even intend to do so, either about the facts or about what he takes the facts to be. What he does necessarily attempt to deceive us about is his enterprise. His only indispensably distinctive characteristic is that in a certain way he misrepresents what he is up to (Frankfurt, 2009, p. 54).

There are two ways of making the argument that naturalistic telemetry could facilitate bullshit detection, the first of which is problematic. Although I don't endorse the argument, it is worth discussing where it goes wrong.

5.3.1 Argument 1

1. Psychologists are obliged to sell the importance of their theories in order to secure publications and grants.
2. To meet their obligations, psychologists market theories as being general in scope without any due regard for whether the theories are actually generalizable.
3. To engage in #2 is to bullshit.

The conclusion to Argument 1 is that psychologists are obliged to bullshit. The first premise is also plausible, as granting agencies and journals do require that researchers speak to the *potential importance* of their theories.

The problem lies with premise #2, which is potentially uncharitable to researchers, and its defense would no doubt entail a heavy burden of proof. Nevertheless, a starting point would be to argue that most theorists who intend their theories to generalize cannot *know* whether their theories will generalize, as a thorough evaluation of a theory's generality is, at the very least, a massive undertaking. Such evidence is unlikely to be present, especially *prior* to funding. The obligation to market one's theory as having a general scope, even when such claims are impossible to support prior to funding and research, might then be seen as a breeding ground for bullshit.

> Bullshit is unavoidable whenever circumstances require someone to talk without knowing what he is talking about. Thus the production of bullshit is stimulated whenever a person's obligations or opportunities to speak about some topic are more excessive than his knowledge of the facts that are relevant to that topic (Frankfurt, 2009, p. 63).

The core of Argument 1, then, is that the challenges inherent in supporting the generality of some theories is so extreme that requiring researchers to speak to theory generalizability is (often) to invite speculation. Insofar as researchers are obliged to testify to the generality of their theory, researchers must bullshit out of necessity and not malice.

The problem with premise # 2 is that researchers can sell the importance of their theories without bullshit. They only need to speak to the generality of their theory while acknowledging their speculation for what it is. This would involve no intent to *misrepresent* their theory as something it is not, and so it would not constitute bullshit.

What is useful about argument 1 is that it reveals the relationship between bullshitting and community expectations. If generality becomes easier to evaluate, and consequently if a reliance on speculation to justify theoretical importance becomes less appropriate, then baseless professions of theory generality will constitute bullshit where the intention was to misrepresent the importance of a theory. This leads me to a revision of argument 1.

5.3.2 Argument 2

1. Psychologists are obliged to sell the importance of their theories in order to secure publications and grants.

2. To meet their obligations, psychologists may market theories as being general in scope, but community standards require that these claims be backed up by evidence.

3. If a theorist makes claims about their theory's generality without due regard for community standards of care or accuracy, their claims of generality are bullshit.

4. As Big Data provides a myriad of novel data sources[10] that allow for examination of questions about theory generality, researchers with a genuine interest in their theory's generality will increasingly seek out useful naturalistic telemetry, provide guidance to those working with such data sources, or will at least go to lengths to explain why their theory cannot be extended to available datasets.

Claim #1 remains unchanged, and Claim #3 follows from the present definition of bullshit. Claim #2 has been modified to refer to community standards of backing up pronouncements of generality with evidence. It is also assumed here that scientific communities change, and the amount of evidence required to support a claim will change with the scientific communities capacity to actually acquire such evidence.

Claim #4 points to the possibility that naturalistic telemetry might allow the scientific community to acquire more evidence pertaining to theory generality, which would impact the quality of evidence required to support claims of generality. It rests, of course, on the empirical premise that researchers will actually be able to find naturalistic telemetry that speak to theory generalizability. The premise is plausible given the recent use of naturally occurring datasets to test theory (Goldstone & Lupyan, 2016; Huang et al., 2017; Stafford & Haasnoot, 2017; Thompson et al., 2014; Thompson, McColeman, et al., 2017).

In short, if naturalistic telemetry becomes the theory evaluation tool that it could be, then the availability of data sources to test theory generality could change the expectations of due care that are placed upon researchers. In this case, professions of theory generality which are bullshit could be identified by professors explicit attempts to test their theories generality, to specify data sources where such tests will be feasible, or to clarify what sort of naturally occurring dataset would be required to examine generalizability.

One can sensibly argue that, at least from the purview of argument 2, naturalistic telemetry is not a bullshit detector, but that the development of theory evaluation tools changes community standards. Statements which previously were mutually understood to be speculative would, after emergence of data sources become bullshit. However, while argument 2 may not imply that naturalistic telemetry is a bullshit detector, it certainly implies that emerging data sources could have a civilizing effect on science that makes bullshit detection possible.

More can be done to clarify how emerging data sources could actually be used to improve tests of theory evaluation. From previous work using naturally occurring datasets (Goldstone & Lupyan, 2016; Huang et al., 2017; Stafford & Haasnoot, 2017; Thompson et al., 2013, 2014; Thompson, McColeman, et al., 2017), and from the studies reported in Chapters

[10]Importantly, by *novel data sources* I am not simply referring to larger quantities of data. I take it to also imply the acquisition of data on an increasing variety of complex human behaviours. The availability of detailed data on naturally occurring video game performance is one example. On this understanding, we will have access to an increasing variety of data sources as increasing varieties of human behaviour become computer mediated.

Three and Four, it is possible to distinguish between two non-exclusive strategies for using naturalistic telemetry to test theory generalizability in complex skill learning. I call these the model-organism approach and the free-for-all (FFA) approach.

5.4 The model-organism approach: A traditional strategy for using convenient data sources

In 1973, Simon and Chase argued that Chess could serve as the drosophila of cognitive science.

> As genetics needs its model organisms, its Drosophila and Neurospora, so psychology needs standard task environments around which knowledge and understanding can cumulate (Simon & Chase, 1973).

'Cumulation', here, refers to an integration of research. Choosing a single complex task in which researchers can pool resources and build up a comprehensive understanding of performance is supposed to help researchers avert the state of affairs where increasingly specialized scientific theories cannot be integrated into a comprehensive theory (Newell, 1973). Of course, one major difference between studying drosophila and Chess players is that the latter are much more expensive to produce and study in quantity. Conveniently, one of the largest changes of the Big Data era is the appearance of many new task environments from which data collection is easier and data sources richer.

One might then propose, as Thompson et al. (2013) did, that the study of skill learning transition to model task environments in which telemetry are available. This constitutes what I call the *model-organism approach* to the exploitation of emerging data sources in psychology. According to this strategy, psychologists ought to survey new data sources for domains of theoretical interest and methodological value. Some of the reasons for studying StarCraft 2, for example, are both theoretical and practical (e.g., StarCraft 2 imposes substantial motor demands which are of theoretical interest) and others methodological (e.g., availability of longitudinal data, availability of demographic information; also see Chapter Two).

Once rich new data sources have been surveyed and a suitable model-organism established, researchers should target the domain with every method available. Hypotheses grounded in quantitative naturalistic telemetry can be tested in the laboratory, laboratory predictions can be examined against natural performance, and where telemetry allows for

the reproduction of the performance, important qualitative observations can be much more easily shared between researchers.[11]

Inspiration for the present work draws heavily upon the model-organism approach. Indeed, the goal of Chapter Three was to lay the ground work for future utilization of StarCraft 2 as a model task domain. This is why it is appropriate for grant-funded researchers to invest effort into modelling the differences between 'Zerg' and 'Protoss' response times. Such a foundation would undoubtedly be required for future researchers to understand how psychological theories generalize to StarCraft 2. One might adopt a variety of methods in the hopes of studying, for example, how bottom up processes might influence oculomotor attentional allocation in StarCraft 2. However, results from such work would have to be interpreted in light of gameplay race, and differences in how gameplay history impacts the speed of actions following screen shifts.

Chapter Four also draws upon the model-organism approach insofar as it aims to unpack and clarify the nature of age-related differences found in previous studies (Thompson et al., 2014). While it would be legitimate to attempt and identify age-related differences in other performance domains (e.g., Tekofsky et al., 2015), the model-organism approach would encourage the systematic unpacking of age-related differences in a single domain. Ideally, such would eventually be fully integrated with laboratory findings on age-related differences.

The model-organism approach is likely to be the least contentious use of telemetry for the study of skill learning. Despite its methodological advantages, it can be seen as a simple extension of expertise research into new domains. For a picture of how it fits into cognitive psychology more broadly, one need only turn to the likes of Newell (1973). The hope is that laboratory studies of basic processes could eventually be integrated into the explanation of a complex skill by probing the phenomenon with a wide range of methods. However, while the model-organism approach to exploiting naturalistic telemetry fits relatively well within traditional cognitive psychology, one objection deserves mention.

5.5 The Streetlight Effect: an Objection to the Model-Organism Approach

One objection to the model-organism approach is that it resembles the Drunkard's search (Kaplan, n.d., p. 11), a fallacy whereby one conducts inquiry solely where data are freely

[11]Obviously not all aspects of the performance context are captured in StarCraft 2 Replay files. I do not know about the distractions in a specific player's home environment, for example. Nevertheless, there is methodological value in the capacity to replay performance with as much of the game context as would have been available to anyone watching the game from home. It should be noted, however, that researchers still need to take steps to preserve qualitative observations for later researchers. For example, older replay files can no longer be viewed using the recent versions of StarCraft 2.

available (e.g., the streetlight) rather than where theory dictates I should look.[12] This strategy can result in obviously silly behaviour (e.g., if I look for my keys under the streetlight where it is bright rather than the dark places where I remember dropping them). This concern is sometimes leveled against the use of convenient data sources found in the Big Data era (Raffaelli et al., 2014).[13]

Of course, one possible response to this objection is to argue that current psychological theory is a poor guide. Perhaps our theories are too underdeveloped to provide direction for future work. The analogy between psychological science and the man searching for his keys breaks down, therefore, because the man with the keys has good reason to think the keys are *not* under the streetlight.

However, one need not bite the bullet and insist that psychological theory is a useless guide for future work. A much less contentious response to the streetlight problem is that, in the history of science, model-organisms are selected on theoretical grounds as well as methodological grounds. In 1973, one of the crucial methodological advantages of using *Caenorhabditis elegans* as a model-organism was the capacity to examine the animal's nervous system more directly.

> One experimental approach to these problems is to investigate the effects of mutations on nervous systems. In principle, it should be possible to dissect the genetic specification of a nervous system in much the same way as was done for biosynthetic pathways in bacteria or for bacteriophage assembly. However, one surmises that genetical analysis alone would have provided only a very general picture of the organization of those processes. [...] Behavior is the result of a complex and ill-understood set of computations performed by nervous systems and it seems essential to decompose the problem into two: one concerned with the question of the genetic specification of nervous systems and the other with the way nervous systems work to produce behavior. Both require that we must have some way of analyzing the structure of a nervous system (Brenner, 1974) (p. 72).

A key advantage of the *Caenorhabditis elegans*, therefore, was the availability of methods which allowed analysis of both how genetics contributed to the development of the animal's nervous system and how the relatively simple nervous system produced behaviour.

[12]Thanks goes to Jeremy Carpendale for pointing this out.

[13] It is important to not conflate *methodological considerations in choosing a research domains* and *convenience sampling* here. Telemetry research need not use convenience sampling. Indeed, Huang et al. (2017) had access to census data from the 2010 game Halo Reach and so they chose Halo Reach as their domain of study. They could have decided, instead, to scrounge up data on the 2001 game Halo: Combat Evolved, though the age of the game makes it possible that valuable data would be unavailable or lost. Huang and colleagues use methodological considerations in choosing their research domain, but they did not use *convenience sampling* from their domain of interest (a particular game mode of Halo Reach performance).

A model-organism with a relatively simple nervous system was therefore ideal. This choice of methodological convenience, of course, risked running into problems with the streetlight effect. The questions of most central importance about the relationship between genetics and behaviour might only be answered by turning to animals with more complex nervous systems.

To forbid any methodological considerations from the selection of a model-organism would probably have denied biology at least some of its most useful model-organisms.[14] Consequently, we should allow psychologists to use methodological considerations in the selection of their task environments, particularly when there are theoretical reasons that those data are of interest. In the present case, a complete theory of skilled performance would have to apply as much to StarCraft 2 as it would to chess, so the choice of StarCraft 2 as a model-organism remains defensible.

5.6 The FFA Approach: A more Aggressive Alternative to the Model-Organism Approach

The model-organism approach does encourage the use of emerging data sources, and in this sense it would likely bring psychology closer to embracing naturalistic telemetry. However, it is important to note that it is also a very *conservative* approach to the utilization of emerging data sources. The basic tenant of the model-organism approach is for psychology to colonize a few rich pastures. This might qualify as Big Data today only because many of these data sources remain suitably unwieldy. But once the move towards the new data sources has been made, there is no requirement that psychology continue to change with the times, and no guarantee that such data will always qualify as Big. It is entirely in keeping with the model-organism approach, for example, that psychology keep its special focus on Real-Time Strategy Video Games like StarCraft 2 as long as the genre continues to exist. In other words, the model-organism approach may only exploit emerging data sources when it is seeking out new model-organisms. Once model-organisms have been found and standard analysis protocols established, researchers adopting the *model-organism approach* will need good methodological reason to tackle new domains or the challenges of larger datasets.

Due to this conservatism, a feature of the model-organism approach is that the list of well-studied domains will remain somewhat unexplored. This is potentially problematic in expertise research where, unlike the relative homogeneity in the behavior of genetics across biological species, different tasks might have entirely different psychological requirements. Outcomes of simple battles in StarCraft 2, for example, are deterministic and predictable. Skilled players can recite how many attacks it will take for a 'Roach' to kill a 'Zergling'

[14]To get a sense of the value of *Caenorhabditis elegans* to biological research, see Brenner's Nobel Lecture entitled 'Nature's gift to Science' (Brenner, 2002).

(Liquid, n.d.). StarCraft 2 is probably a poor choice, therefore, for studying skill in domains requiring probabilistic reasoning. Similarly, the standard of skill in StarCraft 2 are 1v1 player games, making it less ideal for studies of team coordination. It is entirely possible that animal models of human biology will be much more valuable than video-game models of other forms of expertise. If so, the model-organism approach might just chain psychology to a few bad models which say little about the domains of skill that we really care about (e.g., medicine).

An alternative approach, which I will call the Free-For-All (FFA) approach, is to take full advantage of naturalistic telemetry. It proceeds by simply turning to data on any complex behaviour that is relevant to the theories under consideration, even when these data sources come from previously unstudied domains. Given that the understanding of behaviour in StarCraft 2, which has only been examined in detail in a handful of papers (Huang et al., 2017; Lewis et al., 2011; Thompson et al., 2013, 2014; Thompson, McColeman, et al., 2017), is still in its infancy, I could have construed Thompson, Blair, Chen, and Henrey's (2014) search for age-related differences in StarCraft 2 play as fitting with the FFA approach.

I mention two caveats of the FFA approach here. First, it is important that, while the FFA approach relies on the relatively free exploitation of naturalistic telemetry from novel domains of behaviour, it relies on replication as much as, and perhaps more, than traditional methods. The lack of domain knowledge of an FFA researcher means that apparent disconfirmations of theory generalizability could be illusory or artifactual. While I am still unable to explain the strong learning effects associated with 2v2 performance, attempts to probe this effect revealed that the value of off-race experience might be substantively different for non-dedicated and dedicated individuals. The ad hoc removal of dedicated players therefore leaves one with data that would *fit* a preexisting hypothesis that off-race experience should transfer. The fact that this finding was ad hoc makes it unclear whether I should express any confidence in the hypothesis that off-race experience will transfer to dominant-race experience. Similarly, Thompson, McColeman, et al. (2017) found that First Action Latencies are slower in fixations where more actions are being performed, as predicted by the view that such sequences need to be loaded into a motor-buffer prior to execution (Henry & Rogers, 1960; Klapp & Jagacinski, 2011), but only after excluding fixations containing one action, which appeared sufficiently delayed to be unrepresentative. Consequently, the FFA approach, because it does not target a single domain, is unlikely to lead to clear confirmations and disconfirmations of theory. The availability of data means that theorists will have ample opportunity to initiate an ad hoc rescue of their theory. So while the FFA approach involves the free adoption of new domains as needed, replication is nevertheless crucial for highlighting the inadequacy of a particular theory. Importantly, I do not take this continued reliance on replication as a weakness of the FFA approach, as replication remains crucial to all scientific methods.

The second caveat is that the FFA approach will sometimes be less susceptible to the streetlight-effect than the model-organism approach because, while the FFA approach does take advantage of convenient data sources, it is free to choose from a wider array of model-organisms. If a variety of data sources are available that are both theoretically appropriate and methodologically convenient, there seems to be little reason to complain that scientists are only 'searching where the lights are'.

However, the FFA approach has a problem if it is expected to establish construct validity of its measures as corresponding to the psychological entities found in the lab. For one, establishing construct validity of a measure is frought with controversy, especially in a complex task.[15] All complex tasks will draw on a variety of psychological abilities to varying degrees and in varying fashions. StarCraft 2, for example, is too complex for mean FAL not to require some contribution of perceptual speed, decision making speed, and psychomotor speed. The model-organism approach deals with this problem by targeting a single domain with a variety of methods in hope that we will eventually be able to tease apart the factors contributing to variables such as median FAL. The model-organism approach can then speak to laboratory theory by invoking a well established story of how constructs cumulate in complex behaviour. The model-organism approach eventually overcomes problems with measurement construct validity by first growing the roots of the nomological network of scientific laws that, according to Cronbach and Meehl, define our laboratory constructs. FFA researchers, like other researchers studying a new domain, will have no such recourse, and so they will have trouble convincing theorists that any of their variables correspond to laboratory measures, making it difficult, at least in some research circles, to establish that their work has any *theoretical* interest.

Braun (2015) recognizes the problem of measurement validity in the context of Big Data, and propose that Big Data researchers will have to adopt different methods and standards of validation. For example, Braun and Kuljanin point out that, since naturalistic telemetry won't contain any of psychology's established measures, the standard practice of demonstrating measurement validity by pointing to correlations involving established measures will be difficult to follow and, even where this effort is made, there are a number of statistical complications making standard practice problematic (Braun, 2015).

For my purposes, it is only important to make two points about establishing construct validity in the context of naturalistic telemetry. First, while revising our framework for validating measures may allow for better Big Data science, as Braun (2015) propose, nat-

[15]There is a sense in which demonstrating construct validity is always difficult. In fact, construct validity is also never complete according ot Cronbach and Meehl (1955). In complex task environments, attempting to validate a construct is particularily demanding, as the complexity of behaviours under discussion will make proposals of construct validity highly contentious. For example, the claim that RCSpeed is a valid measure of psychomotor speed will be especially controversial because it is easy to speculate about higher-level cognitive factors that could impact the right-clicking behavior of a participant.

uralistic telelmetry can only fulfill its promise as a tool for examining theory generality if *theorists* agree to the change.

Secondly, small revisions to construct validity theory will likely not do. The measures taken on an FFA approach are doomed to linger at the edges of our nomological network. That is part of the FFA strategy. Therefore, if one truly believes that the nomological network bestows our measures with meaning, as Cronbach and Meehl (1955) thought, then the meaningfulness of FFA telemetry data must literally be unclear. Consequently, researchers should not adopt a reading of construct validity lightly, as such decisions shape methods in profound ways.

5.7 Conclusion

I have found that, at least in the case of StarCraft 2, the broader research strategy motivating the exploitation of new data sources can shape the methodological and conceptual difficulties one encounters. Insofar as Chapter Three is motivated by the model-organism approach (e.g., it is seen as a necessary precursor to longitudinal work in the study of this task domain), it fits relatively cleanly within extant methods and philosophy of science. A primary contribution of Chapter Four is an attempt to test the generalizability of processing speed theory through a novel data source, so it seems to fit more cleanly with a FFA strategy.

Of course, the model-organism approach and the FFA approach are not mutually exhaustive or exclusive. Actual research practice is likely to adopt some amalgam of these research strategies. Chapter Four, for example, was largely framed as taking an FFA approach insofar as it was a single study introducing possible psychomotor speed measures that could be used to test a prediction of processing speed theory. It could have also been framed as taking a model-organism approach if I instead emphasized how these results will lay the foundation for future work in StarCraft 2. For example, if the goal is to see how basic cognitive capacities cumulate in a complex task, then Chapter Four reveals something about how skill is preserved across aging in a few different performance indicators. I found, for example, evidence that differences in median FAL are not due solely to individual differences in motor dexterity (which presumably is more closely tracked by finger-tapping), as median FAL and RCSpeed are only weakly related. Such findings will no doubt be valuable to future work in this domain.

Furthermore, there have been cases where the two strategies provided the same recommendations. Frequently studied model-organisms are sometimes convenient to study, though a proper historical analysis of this issue is beyond my scope. Brenner (1974), in defending his choice of Caenorhabditis elegans was obliged to acknowledge that prior work made the study of drosophila very convenient for studying genetics, but defended a place for Caenorhabditis elegans on the grounds that science required an organism with a simpler

nervous system. Even today, the study of StarCraft 2 is more convenient thanks to the existence of at least seven published studies (Huang et al., 2017; Lewis et al., 2011; Thompson et al., 2013, 2014; Thompson, Leung, Blair, & Taboada, 2017; Thompson, McColeman, et al., 2017; Yan & Cheung, 2015). However, the availability of emerging data sources, and otherwise private data sources made available through industry, will probably lead to more contexts where model-organism approach and the FFA approach differ. In deciding upon whether to embrace naturalistic telemetry, psychology would do well to consider the advantages and limitations of both strategies, and the deeper conceptual complications that their adoption entails.

References

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, *16*(07), 1–2. Retrieved from `https://www.wired.com/2008/06/pb-theory/` doi: 10.1016/j.ecolmodel.2009.09.008

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408.

Andress, G. (n.d.). *MLG StarCraft 2: 2v2 Tips and Strategies.* Retrieved 2018-01-19, from `http://www.majorleaguegaming.com/news/mlg-starcraft-2-2v2-tips-and-strategies`

Austin, D., Jimison, H., Hayes, T., Mattek, N., Kaye, J., & Pavel, M. (2011). Measuring motor speed through typing: a surrogate for the finger tapping test. *Behavior research methods*, *43*, 903–9. doi: 10.3758/s13428-011-0100-1

Barnett, A. G., Pols, J. C. V. D., & Dobson, A. J. (2005). Regression to the mean : what it is and how to deal with it. *International Journal of Epide*, *34*(1), 215–220. doi: 10.1093/ije/dyh299

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R.* Retrieved from `http://lme4.r-forge.r-project.org/book/`

Battle.net Leagues. (n.d.). Retrieved 2017-07-08, from `http://wiki.teamliquid.net/starcraft2/Battle.net_Leagues`

Batzokis, G., Lu, P. H., Tingus, K., Mendez, M. F., Richard, A., Peters, D. G., ... Mintz, J. (2011). Lifespan trajectory of myelin integrity and maximum motor speed. *Neurobiology of Aging*, *31*(9), 1554–1562. doi: 10.1016/j.neurobiolaging.2008.08.015.Lifespan

Belicza, C. (2014). *SC2Gears.* Retrieved from `https://sites.google.com/site/sc2gears/`

Berthelot, G., Len, S., Hellard, P., Tafflet, M., Guillaume, M., Vollmer, J.-c., ... Toussaint, J.-f. (2012). Exponential growth combined with exponential decline explains lifetime performance evolution in individual and human species. *Age*, *34*, 1001–1009. doi: 10.1007/s11357-011-9274-9

Birk Diedenhofen. (2016). Package 'cocor': Comparing Correlations. Retrieved from `http://comparingcorrelations.org/`

Bollier, D. (2010). *The Promise and Peril of Big Data.* The Aspen Institute. doi: 10.5210/fm.v19i6.5331

Bosman, E. a. (1993). Age-related differences in the motoric aspects of transcription typing skill. *Psychology and aging*, *8*(1), 87–102. doi: 10.1037/0882-7974.8.1.87

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication Society*, *15*(5), 1–5. doi: 10.1126/science.1243089

Braun, M. T. (2015). Big Data and the Challenge of Construct Validity. *Industrial and Organizational Psychology*, *8*(4), 521–527.

Brenner, S. (1974). Caenorhabditis elegans. *Genetics*, *77*, 71–94.

Brenner, S. (2002). Nature's Gift to Science. In P. K. Grandin (Ed.), *Les prix nobel.* Nobel Foundation.

Bryan, W. L., & Harter, N. (1897). Studies in the physiology and psychology of telegraphic language. *The psychological review*, *4*(1), 27–53.

Bryan, W. L., & Harter, N. (1899). Studies on the telegraphic language. The aquisition of a hierarchy of habits. *The psychological review*, *4*(4), 345–375.

Canty, A., & Ripley, B. D. (2017). boot: Bootstrap R (S-Plus) Functions [Computer software manual].

Cerella, J. (1990). Aging and information processing rate. In *Behaviour, aging, and the nervous system* (pp. 201–222). San Diego: Academic Press.

Colcombe, S., & Kramer, A. F. (2003). Fitness Effects on the Cognitive Function of Older Adults : A Meta-Analytic Study. *Psychological Science*, *14*(2), 125–130.

Cook's Distance. (n.d.). Retrieved 2018-01-20, from `https://www.mathworks.com/help/stats/cooks-distance.html?s_tid=gn_loc_drop`

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi: 10.1037/h0040957

Curran, J., & Chang, D. (2012). dafs: Data analysis for forensic scientists [Computer software manual]. Retrieved from `https://cran.r-project.org/package=dafs`

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications.* Cambridge: Cambridge University Press. Retrieved from `http://statwww.epfl.ch/davison/BMA/`

Donner, Y., & Hardy, J. L. (2015). Piecewise power laws in individual learning curves. *Psychonomic Bulletin Review*, *22*(5), 1308–1319. Retrieved from `http://link.springer.com/article/10.3758/s13423-015-0811-x` doi: 10.3758/s13423-015-0811-x

Douglas, A., Bolker, B., Walker, S., Singmann, H., Dai, B., & Scheipl, F. (2017). *Package 'lme4'.* Retrieved from `https://cran.r-project.org/web/packages/lme4/lme4.pdf`

Ebbinghaus, H. (1885). *Über das Gedächtnis: untersuchungen zur experimentellen psychologie.* Leipzig: Duncker Humblot.

Ericsson, K. A., Krampe, R. T., & Tesch-roemer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. , *100*(3), 363. doi: 10.1037//0033-295X.100.3.363

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual review of psychology*, *47*(1), 273–305. doi: 10.1146/annurev.psych.47.1.273

Esports world championships at Blizzcon 2016. (n.d.). Retrieved 2017-10-05, from `https://blizzcon.com/en-us/news/20089729/esports-world-championships-at-blizzcon-2016`

Essential facts about the computer and video game industry (Tech. Rep.). (2017). Entertainment Software Association. Retrieved from `http://www.theesa.com/wp-content/uploads/2017/06/!EF2017_Design_FinalDigital.pdf`

Feltovich, P. J., Prietula, M. J., & Ericsson, K. A. (2006). Studies of Expertise from Psychological Perspectives. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman

(Eds.), *Cambridge handbook of expertise and expert performance* (pp. 41–67). Cambridge, MA: Cambridge University Press.

Fitts, P. M. (1954). The Information Capacity of the Human Motor system in controlling the amplitude of movement. *Journal of Experimental Biology*, *47*(6), 381–391. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13174710 doi: 10.1037/h0055392

Fodor, J. (2008). *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Fozard, J. L., Vercruyssen, M., Reynolds, S. L., Hancock, P. A., & Quilter, R. E. (1994). Age differences and changes in reaction time: The Baltimore longitudinal study of aging. *Journal of Gerontology*, *49*(4), P179–P189. Retrieved from http://geronj.oxfordjournals.org/cgi/doi/10.1093/geronj/49.4.P179 doi: 10.1093/geronj/49.4.P179

Frankfurt, H. G. (2009). *On bullshit*. Princeton, New Jersey: Princeton University Press. doi: 10.1080/10584600701641920

Goldman, W. P., Baty, J. D., Buckles, V. D., Sahrmann, S., & Morris, J. C. (1999). Motor dysfunction in mildly demented AD individuals without extrapyramidal signs. *Neurology*, *53*(5), 956–962. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10496252 doi: 10.1212/WNL.53.5.956

Goldstone, R. L., & Lupyan, G. (2016). Discovering Psychological Principles by Mining Naturally Occurring Data Sets. *Topics in Cognitive Science*, *8*(3), 548–568. doi: 10.1111/tops.12212

Gray, W. D., & Lindstedt, J. K. (2016). Plateaus, Dips, and Leaps: Where to Look for Inventions and Discoveries During Skilled Performance. *Cognitive Science*, *41*(7), 1–33. doi: 10.1111/cogs.12412

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534–537. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12774121 doi: 10.1038/nature01647

Green, C. S., Kattner, F., Eichenbaum, A., Bediou, B., Adams, D. M., Mayer, R. E., & Bavelier, D. (2017). Playing Some Video Games but Not Others Is Related to Cognitive Abilities. *Psychological Science*, *28*(5), 095679761664483. Retrieved from http://journals.sagepub.com/doi/10.1177/0956797616644837 doi: 10.1177/0956797616644837

Henry, F. M., & Rogers, D. E. (1960). Increased response latency for complicated movements and a "memory drum" theory of neuromotor reaction. *Research Quarterly. American Association for Health, Physical Education and Recreation*, *31*(3), 448–458. APA.

Herbrich, R., Minka, T., & Graepel, T. (2007). TrueSkill^{TM}: A Bayesian Skill Rating System. *Advances in Neural Information Processing Systems*, *20*, 569–576.

Højsgaard, S., & Halekoh, U. (2016). doBy: Groupwise Statistics, LSmeans, Linear Contrasts, Utilities [Computer software manual]. Retrieved from https://cran.r-project.org/package=doBy

Hope, R. M. (2013). Rmisc: Rmisc: Ryan Miscellaneous [Computer software manual]. Retrieved from https://cran.r-project.org/package=Rmisc

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363.

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

Internet Live Stats: Twitter Usage Statistics. (n.d.). Retrieved 2017-10-12, from http://www.internetlivestats.com/twitter-statistics/

Kaplan, A. (n.d.). *The Conduct of Inquiry: Methodology for Behavioral Science* (Fourth ed.). New Brunswick, NJ: Transaction Publishers.

Kari, T., & Karhulahti, V.-M. (2016). Do E-Athletes Move? *International Journal of Gaming and Computer-Mediated Simulations*, *8*(4), 53–66. Retrieved from http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJGCMS.2016100104 doi: 10.4018/IJGCMS.2016100104

Keetch, K. M., Lee, T. D., & Schmidt, R. A. (2008). Especial Skills: Specificity Embedded Within Generality. *Journal of Sport Exercise Psychology*, *30*, 723–736.

Keller, F. S. (1958). The phantom plateau. *Journal of the experimental analysis of behavior*, *1*(1), 1–13. doi: 10.2466/pms.1985.61.1.55

Kendall. (1945). *Advanced Theory Of Statistics Vol-I.* London: Griffin.

Khazaal, Y., van Singer, M., Chatton, A., Achab, S., Zullino, D., Rothen, S., ... Thorens, G. (2014). Does Self-Selection Affect Samples' Representativeness in Online Surveys? An Investigation in Online Video Game Research Yasser. *Journal of Medical Internet Research*, *16*(7), 1–10. doi: 10.2196/jmir.2759

Kim, S. (2015). ppcor: Partial and Semi-Partial (Part) Correlation [Computer software manual]. Retrieved from https://cran.r-project.org/package=ppcor

Kitchin, R. (2014). Big Data , new epistemologies and paradigm shifts. *Big Data Society*(June), 1–12. doi: 10.1177/2053951714528481

Klapp, S. T., & Jagacinski, R. J. (2011). Gestalt principles in the control of motor action. *Psychological bulletin*, *137*(3), 443–462. doi: 10.1037/a0022361

Lane, D. M. (n.d.). *Influential Observations.* Retrieved 2018-01-20, from http://onlinestatbook.com/2/regression/influential.html

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *Gartner*.

Lewis, J. M., Trinh, P., & Kirsh, D. (2011). A Corpus Analysis of Strategy Video Game Play in Starcraft : Brood War. *Proceedings of the Annual Meeting of the Cognitive Science*, *33*(33), 687–692.

Lezak, M. (2004). *Neuropsychological Assessment.* New York, NY: Oxford University Press.

Lin, M., & Lucas, H. C. (2013). Too Big to Fail : Large Samples and the p -Value Problem. *Information Systems Research*, *7047*(August 2016), 1–12. doi: http://dx.doi.org/10.1287/isre.2013.0480

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., & Ye, K. Q. (2006). Variable Selection for Gaussian Process Models in Computer Experiments. *Technometrics*, *48*(4), 478–490.

Liquid, T. (n.d.). *Zergling (Legacy of the Void).* Retrieved 2017-09-29, from http://wiki.teamliquid.net/starcraft2/Zergling_(Legacy_of_the_Void)

Loehr, J. D., & Palmer, C. (2007). Cognitive and biomechanical influences in pianists' finger tapping. *Experimental Brain Research*, *178*, 518–528. doi: 10.1007/s00221-006-0760-8

Mangiafico, S. S. (2016). Confidence Intervals for Medians. In *Summary and analysis of extension program evaluation in r.* New Brunswick, NJ.: Rutgers Cooperative Extension. Retrieved from http://rcompanion.org/handbook/E_04.html

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hypothe-*

ses, *82*(4), 405–411. Retrieved from http://www.sciencedirect.com/science/article/pii/S0306987713005598 doi: https://doi.org/10.1016/j.mehy.2013.11.030

Mazzocchi, F. (2015). Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO reports*, *16*(10), 1250–1255. doi: 10.15252/embr.201541001

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [Computer software manual]. Retrieved from https://cran.r-project.org/package=e1071

Miller, E. M. (1994). Intelligence and Brain Myelination: Hypothesis. *Personality and Individual Differences*, *17*(6), 803–832.

Morrow, D. G., Leirer, V. O., & Altieri, P. A. (1992). *Aging, expertise, and narrative processing.* (Vol. 7) (No. 3). Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0882-7974.7.3.376 doi: 10.1037/0882-7974.7.3.376

Moxley, J. H., & Charness, N. (2013). Meta-analysis of age and skill effects on recalling chess positions and selecting the best move. *Psychonomic bulletin review*, *20*, 1017–1022. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23508364 doi: 10.3758/s13423-013-0420-5

Muggeo, V. M. R. (2008). segmented: An R package to Fit Regression Models with Broken-Line Relationships. *R News*, *8*(1), 20–25. Retrieved from http://cran.r-project.org/doc/Rnews/ doi: 10.1159/000323281

Muggeo, V. M. R. (2016). Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling. *Journal of Statistical Computation and Simulation*, *86*(15), 3059–3067. Retrieved from http://www.tandfonline.com/doi/full/10.1080/00949655.2016.1149855 doi: 10.1080/00949655.2016.1149855

Muggeo, V. M. R. (2017). *Segmented : Regression Models with Break-Points/Change-Points Estimation.* Retrieved 2017-08-22, from https://cran.r-project.org/web/packages/segmented/segmented.pdf

Navarro, D. (2015). Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5) [Computer software manual]. Adelaide, Australia. Retrieved from http://ua.edu.au/ccs/teaching/lsr

Newell, A. (1973). You can ' t play 20 questions with nature and win : projective comments on the papers of this symposium. In W. Chase (Ed.), *Visual information processing.* New York: Academic Press.

Ooms, J., James, D., DebRoy, S., Wickham, H., & Horner, J. (2017a). RMySQL: Database Interface and 'MySQL' Driver for R [Computer software manual]. Retrieved from https://cran.r-project.org/package=RMySQL

Ooms, J., James, D., DebRoy, S., Wickham, H., & Horner, J. (2017b). RMySQL: Database Interface and 'MySQL' Driver for R [Computer software manual]. Retrieved from https://cran.r-project.org/package=RMySQL

Pardoe, D. I. (n.d.). *11.2 - Using Leverages to Help Identify Extreme x Values.* Retrieved 2018-01-20, from https://onlinecourses.science.psu.edu/stat501/node/338

Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods*, 1–9. Retrieved from http://dx.doi.org/10.3758/s13428-017-0874-x doi: 10.3758/s13428-017-0874-x

Pearson, K., & Filon, L. . N. . G. . (1898). Mathematical Contributions to the Theory of Evolution . IV . On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *191*, 229–311.

Povel, D. J., & Collard, R. (1982). Structural factors in patterned finger tapping. *Acta Psychologica*, *52*(1-2), 107–123. doi: 10.1016/0001-6918(82)90029-4

R Core Team. (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.*

Raffaelli, D., Bullock, J. M., Cinderby, S., Durance, I., Emmett, B., Harris, J., ... White, P. C. (2014). Big Data and Ecosystem Research Programmes. In G. Woodward, A. J. Drumbrell, D. J. Baird, & M. Hajibabaei (Eds.), *Advances in ecological research: Big data in ecology.* London: Academic Press.

Ranked For Teh Win. (n.d.). Retrieved 2017-08-14, from `http://www.rankedftw.com/`

Redick, T. S., Unsworth, N., Kane, M. J., & Hambrick, D. Z. (2017). Don't Shoot the Messenger: Still No Evidence That Video-Game Experience Is Related to Cognitive Abilities-A Reply to Green et al. (2017). *Psychological science*, *28*(5), 683–686. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/28346108%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5443405` doi: 10.1177/0956797617698527

Rienhoff, R., Hopwood, M. J., Fischer, L., Strauss, B., Baker, J., & Schorer, J. (2013). Transfer of motor and perceptual skills from basketball to darts. *Frontiers in Psychology*, *4*(1), 1–7. doi: 10.3389/fpsyg.2013.00593

Ripley, M. B. (2017). *Package âĂŸboot'.* Retrieved 2017-01-19, from `https://cran.r-project.org/web/packages/boot/boot.pdf`

Rosenbaum, D. A., Kenny, S. B., & Derr, M. A. (1983). Hierarchical Control of Rapid Movement Sequences. , *9*(1), 86–102. doi: 10.1037//0096-1523.9.1.86

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive motor performance. *Cognitive Science*, *6*, 1–36. doi: 10.1016/S0364-0213(82)80004-9

Salthouse, T. A. (1984). Effects of age and skill in typing. *Journal of experimental psychology. General*, *113*(3), 345–371. doi: 10.1037/0096-3445.113.3.345

Salthouse, T. A. (1985). Speed of behavior and its implications for cognition. In J. E. Birren & J. E. Schaie (Eds.), *Handbook of psychology and aging* (pp. 400–426). New York: Van Nostrand Reinhold.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428. Retrieved from `http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.103.3.403` doi: 10.1037/0033-295X.103.3.403

Salthouse, T. A. (1998). Relation of successive percentiles of reaction time distributions to cognitive variables and adult age. *Intelligence*, *26*(2), 153–166. doi: 10.1016/S0160-2896(99)80059-2

Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology*, *54*(1-3), 35–54. doi: 10.1016/S0301-0511(00)00052-1

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, *30*(4), 507–514. doi: 10.1016/j.neurobiolaging.2008.09.023

Salthouse, T. A. (2017). Shared and Unique Influences on Age-Related Cognitive Change. *Neuropsychology*, *31*(1), 11–19. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27808539 doi: 10.1037/neu0000330

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying Fixations and Saccades in Eye-Tracking Protocols. *Proceedings of the Eye Tracking Research and Applications Symposium*, 71–78. doi: 10.1145/355017.355028

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

Sarkar, D., & Andrews, F. (2016). latticeExtra: Extra Graphical Utilities Based on Lattice [Computer software manual]. Retrieved from https://cran.r-project.org/package=latticeExtra

Schaie, K. W. (1965). A General Model for the Study of Development Problems. *Psychological bulletin*, *64*(2), 92–107. doi: 10.1037/h0022371

Schaie, K. W. (2009). "When does age-related cognitive decline begin?" Salthouse again reifies the "cross-sectional fallac". *Neurobiology of Aging*, *30*, 528–529. doi: 10.1016/j.neurobiolaging.2008.12.012

Schmitt, L. (2013). Finger-Tapping Test. In F. R. Volkmar (Ed.), *Encyclopedia of autism spectrum disorders* (p. 1296). New York, NY: Springer New York. Retrieved from https://doi.org/10.1007/978-1-4419-1698-3_343 doi: 10.1007/978-1-4419-1698-3_343

Schroeder, D. H., & Salthouse, T. A. (2004). Age-related effects on cognition between 20 and 50 years of age. *Personality and Individual Differences*, *36*(2), 393–404. doi: 10.1016/S0191-8869(03)00104-1

Simon, H., & Chase, W. (1973). Skill in Chess. *American Scientist*, *61*(August 1973), 392–403.

Simpson, E. . H. . (2017). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society*, *13*(2), 238–241.

Singer, R. (1975). *Motor Learning and Human Performance: An application to physical education skills (2nd edition)*. NY,NY: . Macmillan publishing Co.

Slaney, K. L. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Palgrave Macmillan.

Slaney, K. L., & Racine, T. P. (2013). New Ideas in Psychology What ' s in a name ? Psychology ' s ever evasive construct. *New Ideas in Psychology*, *31*, 4–12. Retrieved from http://dx.doi.org/10.1016/j.newideapsych.2011.02.003 doi: 10.1016/j.newideapsych.2011.02.003

Stafford, T., & Haasnoot, E. (2017). Testing Sleep Consolidation in Skill Learning: A Field Study Using an Online Game. *Topics in Cognitive Science*, *9*(2), 485–496. doi: 10.1111/tops.12232

Standford Encyclopedia of Philosophy. (2003). *Francis Bacon.* Retrieved 2017-12-30, from https://plato.stanford.edu/entries/francis-bacon/#SciMetNovOrgTheInd

Standford Encyclopedia of Philosophy: Measurement in Science. (2015). Retrieved 2017-12-31, from https://plato.stanford.edu/entries/measurement-science/

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (Ed.), *Information processing in motor control and learning* (pp. 117–152). New York: Academic Press.

Team Liquid: Personal Sponsorship. (n.d.). Retrieved 2017-10-05, from http://wiki.teamliquid.net/starcraft2/Personal_Sponsorship

Tekofsky, S., Spronck, P., Goudbeek, M., Plaat, A., & Van Den Herik, J. (2015). Past Our Prime: A Study of Age and Play Style Development in Battlefield 3. *IEEE Transactions on Computational Intelligence and AI in Games*, *7*(3), 292–303. doi: 10.1109/TCIAIG.2015.2393433

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). 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), 1–10. doi: 10.1371/journal.pone.0094215

Thompson, J. J., Leung, B., Blair, M. R., & Taboada, M. (2017). Sentiment analysis of player chat messaging in the video game StarCraft 2 : Extending a lexicon-based model. *Knowledge-Based Systems*, *137*, 149–162. Retrieved from https://doi.org/10.1016/j.knosys.2017.09.022 doi: 10.1016/j.knosys.2017.09.022

Thompson, J. J., McColeman, C. M., Stepanova, E. R., & Blair, M. R. (2017). Using Video Game Telemetry Data to Research Motor Chunking , Action Latencies , and Complex Cognitive-Motor Skill Learning. , *9*, 467–484. doi: 10.1111/tops.12254

Thorndike, E. L. (1908). The Effect of Practice in the Case of a Purely Intellectual Function. *The American Journal of Psychology*, *19*(3), 374–384.

Thorndike, E. L. (1913). *Educational psychology, volume II: The psychology of learning.* New York: Teachers College, Columbia University.

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*, *1*(21), 19–25. Retrieved from http://www.wbc.poznan.pl/Content/325867/5_Trends_Vol21_2014_no1_20.pdf

Torchiano, M. (2017). effsize: Efficient Effect Size Computation [Computer software manual]. Retrieved from https://cran.r-project.org/package=effsize

Tsang, P. S., & Shaner, T. L. (1998). Age, attention, expertise, and time-sharing performance. *Psychology and Aging*, *13*(2), 323–347. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0882-7974.13.2.323 doi: 10.1037/0882-7974.13.2.323

Unsworth, N., Redick, T. S., McMillan, B. D., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2015). Is Playing Video Games Related to Cognitive Abilities? *Psychological Science*, *26*(6), 759–774. Retrieved from http://pss.sagepub.com/cgi/content/abstract/26/6/759 doi: 10.1177/0956797615570367

Verhaeghen, P., Steitz, D. W., Sliwinski, M. J., & Cerella, J. (2003). Aging and dual-task performance: A meta-analysis. *Psychology and Aging*, *18*(3), 443–460. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0882-7974.18.3.443 doi: 10.1037/0882-7974.18.3.443

Video games in europe: Consumer study (Tech. Rep.). (2012). Interactive Software Federation of Europe. Retrieved from http://www.isfe.eu/sites/isfe.eu/files/attachments/euro_summary_-_isfe_consumer_study.pdf

Whitehead, A. L., & Perry, S. L. (2017). Unbuckling the Bible Belt: A State-Level Analysis of Religious Factors and Google Searches for Porn. *The Journal of Sex Research*, *in*

press. Retrieved from http://dx.doi.org/10.1080/00224499.2017.1278736 doi: 10.1080/00224499.2017.1278736

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wittgenstein, L. (1953). Philosophical investigations (G. Anscombe, Trans.).

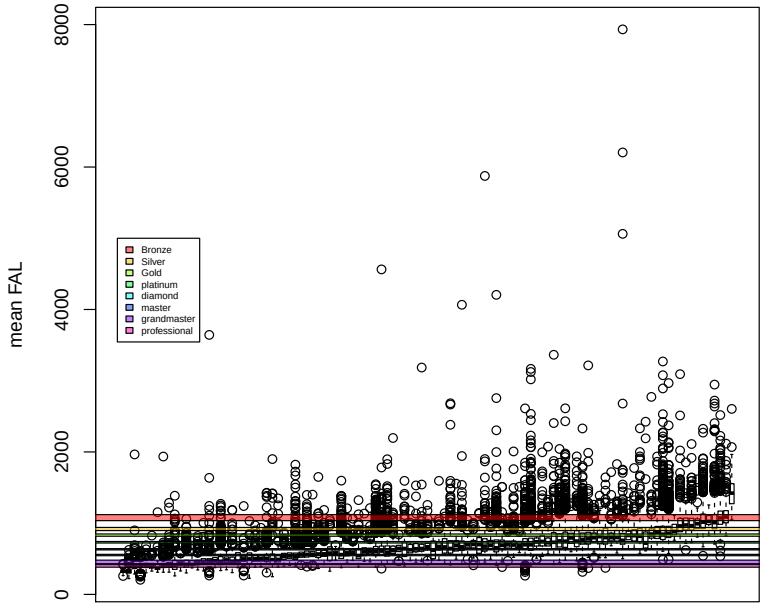
Yamaguchi, M., & Logan, G. (2014). Pushing typists back on the learning curve: Contributions of multiple linguistic units in the acquisition of typing skill. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(6), 1713–1732. doi: 10.1037/xlm0000026

Yamaguchi, M., Logan, G., & Li, V. (2013). Multiple bottlenecks in hierarchical control of action sequences: What does "response selection" select in skilled typewriting? *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 1059–1084. doi: 10.1037/a0030431

Yan, E. Q., & Cheung, G. K. (2015). Masters of Control : Behavioral Patterns of Simultaneous Unit Group Manipulation in StarCraft 2. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 3711–3720). ACM.

Appendix

i Supplementary Figures



Caption:Boxplots of mean FAL for each player in the dataset, with outliers included. Transparent bars are centered around the median *mean FALs* from Thompson, Blair, Chen, & Henrey's (2013) cross-sectional dataset of 3,385 individuals. The width of the bar reflect 2 times the standard error of the median for the league in question (acquired from the standard deviation of 1,000 bootstrapped subsamples, Mangiafico, 2016; Ripley, 2017). The reader should assume that the eight distributions of mean FAL in Thompson et al. (2013) overlap considerably. Finally, these plots are based on a subsample of the dataset with only 82,534 games.

ii Supplementary Results

Statistical analyses were conducted in R using a variety of packages.(D. Bates, Mächler, Bolker, & Walker, 2015; Canty & Ripley, 2017; Curran & Chang, 2012; Davison & Hinkley, 1997; Højsgaard & Halekoh, 2016; Hope, 2013; Hothorn, Bretz, & Westfall, 2008; Kim, 2015; Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2017; Navarro, 2015; Ooms, James, DebRoy, Wickham, & Horner, 2017a, 2017b; Sarkar, 2008; Sarkar & Andrews, 2016; Torchiano, 2017; Wickham, 2009). Unless otherwise specified, the per-test α is 0.05.

ii.1 Models of Experience (Global Analysis)

likelihood ratio tests

```
Data: data
Models:
Model_1: medianfal ~ OneVOnexp + (1 | playerid)
Model_2: medianfal ~ OneVOnexp + Race + (1 | playerid)
        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
Model_1  4 958080 958117 -479036   958072
Model_2  6 956918 956974 -478453   956906 1166.3      2  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Data: data
Models:
Model_2: medianfal ~ OneVOnexp + Race + (1 | playerid)
Model_3: medianfal ~ OneVOnexp + Race + (1 + OneVOnexp | playerid)
        Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
Model_2  6 956918 956974 -478453   956906
Model_3  8 942400 942475 -471192   942384 14522      2  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Data: data
Models:
Model_3: medianfal ~ OneVOnexp + Race + (1 + OneVOnexp | playerid)
Model_3_Pr: medianfal ~ OneVOnexp + Race + Pr_XP + (1 + OneVOnexp | playerid)
           Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
Model_3     8 942400 942475 -471192   942384
Model_3_Pr  9 942379 942463 -471180   942361 23.371      1  1.336e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Data: data
Models:
Model_3_Pr: medianfal ~ OneVOnexp + Race + Pr_XP + (1 + OneVOnexp | playerid)
Model_3_PrTr: medianfal ~ OneVOnexp + Race + Pr_XP + Tr_XP + (1 + OneVOnexp |
Model_3_PrTr:     playerid)
             Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
Model_3_Pr    9 942379 942463 -471180   942361
Model_3_PrTr 10 942366 942460 -471173   942346 14.412      1  0.0001468 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Data: data
Models:
Model_3_PrTr: medianfal ~ OneVOnexp + Race + Pr_XP + Tr_XP + (1 + OneVOnexp |
Model_3_PrTr:     playerid)
model_3_PrTrInteraction: medianfal ~ OneVOnexp + Race + Pr_XP +
Tr_XP + Pr_XP:Race + Tr_XP:Race + model_3_PrTrInteraction:
OneVOnexp:Race + (1 + OneVOnexp | playerid)
                        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
Model_3_PrTr            10 942366 942460 -471173   942346
model_3_PrTrInteraction 16 942255 942405 -471112   942223 122.92      6  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

The presumptive model, $Model_3$

```
Linear mixed model fit by REML ['lmerMod']
Formula: medianfal ~ OneVOnexp + Race + (1 + OneVOnexp | playerid)
   Data: data

REML criterion at convergence: 942366.3

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1952  -0.6123  -0.0311   0.5144  11.4992

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 playerid (Intercept) 25669    160.21
          OneVOnexp    1083     32.92   -0.38
 Residual              5076     71.24
Number of obs: 82768, groups:  playerid, 103

Fixed effects:
            Estimate Std. Error t value
(Intercept)  573.169     15.857   36.15
OneVOnexp    -20.452      3.411   -6.00
RaceTerran    22.061      2.127   10.37
RaceZerg     -42.206      2.147  -19.66

Correlation of Fixed Effects:
           (Intr) OnVOnx RcTrrn
OneVOnexp  -0.366
RaceTerran -0.067  0.003
RaceZerg   -0.074 -0.004  0.573
```

The best model (according to likelihood ratio tests)

```
Linear mixed model fit by REML ['lmerMod']
Formula: medianfal ~ OneVOnexp + Race + Pr_XP + Tr_XP +
Pr_XP:Race + Tr_XP:Race +  OneVOnexp:Race
+ (1 + OneVOnexp | playerid)
   Data: data

REML criterion at convergence: 942191.6

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.2047  -0.6166  -0.0289   0.5178  11.5082

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 playerid (Intercept) 25155    158.60
          OneVOnexp    1097     33.12   -0.40
 Residual              5067     71.18
Number of obs: 82768, groups:  playerid, 103

Fixed effects:
                    Estimate Std. Error t value
(Intercept)          587.8801    15.7851   37.24
OneVOnexp            -20.4279     3.8452   -5.31
RaceTerran             5.8036     3.5786    1.62
RaceZerg             -67.8899     3.4660  -19.59
Pr_XP                  5.8771     3.4595    1.70
Tr_XP                -10.1650     2.8279   -3.59
RaceTerran:Pr_XP       4.5445     0.9227    4.93
RaceZerg:Pr_XP         4.4374     1.0321    4.30
RaceTerran:Tr_XP       1.7813     1.2637    1.41
RaceZerg:Tr_XP         4.3249     1.2331    3.51
OneVOnexp:RaceTerran  -0.3254     0.7056   -0.46
OneVOnexp:RaceZerg     0.4384     0.6468    0.68

Correlation of Fixed Effects:
[Omitted due to space considerations.]
```

ii.2 Models of Experience (Dominant-Race Analysis)

Likelihood ratio tests (dominant-race analysis)

```
************************
Data: Dom_data
Models:
Model_Dom_1: medianfal ~ Dom_xp + (1 | playerid)
Model_Dom_2: medianfal ~ Dom_xp + Race + (1 | playerid)
            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
Model_Dom_1  4 911170 911207 -455581   911162
Model_Dom_2  6 911163 911219 -455576   911151 10.432      2   0.005429 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
************************
Data: Dom_data
Models:
Model_Dom_2: medianfal ~ Dom_xp + Race + (1 | playerid)
Model_Dom_3: medianfal ~ Dom_xp + Race + (1 + Dom_xp | playerid)
            Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
Model_Dom_2  6 911163 911219 -455576   911151
Model_Dom_3  8 897261 897335 -448623   897245 13906      2  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
************************
Data: Dom_data
Models:
Model_Dom_3: medianfal ~ Dom_xp + Race + (1 + Dom_xp | playerid)
Model_Dom_4: medianfal ~ Dom_xp + Race + Off_xp + (1 + Dom_xp | playerid)
            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
Model_Dom_3  8 897261 897335 -448623   897245
Model_Dom_4  9 897259 897342 -448620   897241 4.4554      1    0.03479 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
************************
Data: Dom_data
Models:
Model_Dom_4: medianfal ~ Dom_xp + Race + Off_xp + (1 + Dom_xp | playerid)
Model_Dom_5: medianfal ~ Dom_xp + Race + Off_xp + TwoVTwoXP + (1 + Dom_xp |
Model_Dom_5:     playerid)
            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
Model_Dom_4  9 897259 897342 -448620   897241
Model_Dom_5 10 897030 897122 -448505   897010 231.17      1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
************************
Data: Dom_data
Models:
```

```
Model_Dom_5: medianfal ~ Dom_xp + Race + Off_xp + TwoVTwoXP + (1 + Dom_xp |
Model_Dom_5:     playerid)
Model_Dom_6: medianfal ~ Dom_xp + Race + Off_xp + TwoVTwoXP + N_TwoXP + (1 +
Model_Dom_6:     Dom_xp | playerid)
            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
Model_Dom_5 10 897030 897122 -448505   897010
Model_Dom_6 11 897030 897132 -448504   897008 1.3177      1      0.251
************************
```

The best model according to likelihood ratio tests, $Model_5^{Dom}$ (dominant-race analysis)

```
Linear mixed model fit by REML ['lmerMod']
Formula: medianfal ~ Dom_xp + Race + Off_xp + TwoVTwoXP + (1 + Dom_xp |      playerid)
   Data: Dom_data

REML criterion at convergence: 896975.3

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.4097  -0.6117  -0.0252   0.5160  11.6393

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 playerid (Intercept) 25583.2  159.95
          Dom_xp        752.4   27.43   -0.39
 Residual              4964.5   70.46
Number of obs: 78936, groups:  playerid, 103

Fixed effects:
            Estimate Std. Error t value
(Intercept)  616.731     26.161  23.574
Dom_xp       -14.626      2.899  -5.046
RaceTerran   -48.495     38.773  -1.251
RaceZerg    -102.834     34.101  -3.016
Off_xp         3.416      1.486   2.298
TwoVTwoXP    -24.146      1.582 -15.266

Correlation of Fixed Effects:
           (Intr) Dom_xp RcTrrn RacZrg Off_xp
Dom_xp     -0.222
RaceTerran -0.642  0.003
RaceZerg   -0.728 -0.003  0.492
Off_xp      0.001 -0.033 -0.001 -0.002
TwoVTwoXP  -0.007 -0.092  0.007  0.000 -0.011
```

The presumptive model, $Model_3^{Dom}$ (dominant-race analysis)

```
Linear mixed model fit by REML ['lmerMod']
Formula: medianfal ~ Dom_xp + Race + (1 + Dom_xp | playerid)
   Data: Dom_data

REML criterion at convergence: 897216.1

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.2952  -0.6130  -0.0253   0.5113  11.6122

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 playerid (Intercept) 24852.7  157.65
          Dom_xp        984.7   31.38   -0.37
 Residual              4977.9   70.55
Number of obs: 78936, groups:  playerid, 103

Fixed effects:
            Estimate Std. Error t value
(Intercept)  614.915     25.842  23.795
Dom_xp       -19.288      3.277  -5.885
RaceTerran   -45.096     38.344  -1.176
RaceZerg    -104.480     33.730  -3.098

Correlation of Fixed Effects:
           (Intr) Dom_xp RcTrrn
Dom_xp     -0.219
RaceTerran -0.642  0.004
RaceZerg   -0.729 -0.002  0.492
```

ii.3 Models of Experience (where All Players have Greater than 30 2v2 Games)

$Model_5^{Dom}$

```
Linear mixed model fit by REML ['lmerMod']
Formula: medianfal ~ Dom_xp + Race + TwoVTwoXP + (1 + Dom_xp | playerid)
   Data: Dom_data

REML criterion at convergence: 267970.3

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.0752  -0.6643   0.0347   0.5427   9.3516

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 playerid (Intercept) 23538    153.42
          Dom_xp       1266     35.58   -0.33
 Residual              5295     72.77
Number of obs: 23451, groups:  playerid, 29

Fixed effects:
            Estimate Std. Error t value
(Intercept)  676.717     42.905  15.773
Dom_xp       -14.765      6.884  -2.145
RaceTerran  -146.995     84.071  -1.748
RaceZerg    -132.307     57.974  -2.282
TwoVTwoXP    -23.122      1.674 -13.813

Correlation of Fixed Effects:
           (Intr) Dom_xp RcTrrn RacZrg
Dom_xp     -0.224
RaceTerran -0.488  0.019
RaceZerg   -0.703  0.003  0.358
TwoVTwoXP  -0.005 -0.117 -0.007 -0.004
```

ii.4 Analysis of Age-Related Differences

All of the analyses below contain (1) summary of a linear, non-segmented model, (2) a summary of the segmented model, (3) the result of the Davies test examining whether the effect of age differs between the model segments, and (4) a likelihood ratio test comparing the models described in (1) and (2).

Segmented analysis with log mean FAL as response

```
**********************
******* LogFALMean ********
**********************
****Summary of linear model****

Call:
lm(formula = ResponseVar ~ age * LeagueIdx, data = DATA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75339 -0.13919 -0.00027  0.13861  0.78879

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.7374682  0.0703331  95.794  < 2e-16 ***
age             0.0086715  0.0030080   2.883 0.003967 **
LeagueIdx2     -0.1861560  0.0870972  -2.137 0.032645 *
LeagueIdx3     -0.3158868  0.0820089  -3.852 0.000119 ***
LeagueIdx4     -0.4412528  0.0813510  -5.424 6.25e-08 ***
LeagueIdx5     -0.5259846  0.0833776  -6.308 3.20e-10 ***
LeagueIdx6     -0.8011112  0.0921146  -8.697  < 2e-16 ***
age:LeagueIdx2  0.0015824  0.0037624   0.421 0.674095
age:LeagueIdx3  0.0029437  0.0035393   0.832 0.405622
age:LeagueIdx4  0.0028786  0.0035205   0.818 0.413608
age:LeagueIdx5  0.0001439  0.0036497   0.039 0.968546
age:LeagueIdx6  0.0066820  0.0041406   1.614 0.106673
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 0.214 on 3263 degrees of freedom
Multiple R-squared:  0.4615,Adjusted R-squared:  0.4597
F-statistic: 254.2 on 11 and 3263 DF,  p-value: < 2.2e-16

****Summary of Segmented model****

***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = lm_seg, seg.Z = ~age)
```

```
Estimated Break-Point(s):
   Est. St.Err
24.007  1.483

Meaningful coefficients of the linear terms:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.849884   0.078602  87.146  < 2e-16 ***
age             0.002952   0.003513   0.840   0.4007
LeagueIdx2     -0.196342   0.087036  -2.256   0.0241 *
LeagueIdx3     -0.324987   0.081952  -3.966 7.48e-05 ***
LeagueIdx4     -0.463292   0.081731  -5.669 1.57e-08 ***
LeagueIdx5     -0.573241   0.084849  -6.756 1.67e-11 ***
LeagueIdx6     -0.874381   0.094793  -9.224  < 2e-16 ***
U1.age          0.010261   0.003201   3.205       NA
age:LeagueIdx2  0.002081   0.003760   0.553   0.5800
age:LeagueIdx3  0.003418   0.003538   0.966   0.3341
age:LeagueIdx4  0.004061   0.003543   1.146   0.2519
age:LeagueIdx5  0.002534   0.003733   0.679   0.4973
age:LeagueIdx6  0.010374   0.004289   2.419   0.0156 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 0.2137 on 3261 degrees of freedom
Multiple R-Squared: 0.4633,  Adjusted R-squared: 0.4612

Convergence attained in 2 iterations with relative change 5.831384e-05
****Summary of Davies test****

Davies' test for a change in the slope

data:  formula = ResponseVar ~ age * LeagueIdx ,   method = lm
model = gaussian , link = identity
segmented variable = age
'best' at = 24, n.points = 27, p-value = 0.0122
alternative hypothesis: two.sided

****Summary of Liklihood Ratio test****
comparing models for LogFALMean
Likelihood ratio test

Model 1: ResponseVar ~ age * LeagueIdx
Model 2: ResponseVar ~ age + LeagueIdx + U1.age + psi1.age + age:LeagueIdx
  #Df LogLik Df  Chisq Pr(>Chisq)
1  13 407.71
2  15 413.25  2 11.073    0.00394 **
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Segmented analysis with RCLatency as response

```
**********************
******* RCLatencySegmented ********
**********************
****Summary of linear model****

Call:
lm(formula = ResponseVar ~ age * LeagueIdx, data = DATA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91071 -0.20311 -0.05316  0.12070  2.28706

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.1576155  0.0873693  59.032  < 2e-16 ***
age            0.0244549  0.0039194   6.239 4.96e-10 ***
LeagueIdx      0.0215579  0.0218259   0.988    0.323
age:LeagueIdx -0.0050366  0.0009995  -5.039 4.93e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3411 on 3232 degrees of freedom
Multiple R-squared:  0.1319,Adjusted R-squared:  0.1311
F-statistic: 163.7 on 3 and 3232 DF,  p-value: < 2.2e-16

****Summary of Segmented model****

***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = lm_seg, seg.Z = ~age)

Estimated Break-Point(s):
   Est. St.Err
41.166  9.012

Meaningful coefficients of the linear terms:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.149342   0.088010  58.509  < 2e-16 ***
age            0.024847   0.003952   6.287 3.67e-10 ***
LeagueIdx      0.022488   0.021938   1.025    0.305
U1.age        -0.128474   0.484531  -0.265       NA
age:LeagueIdx -0.005080   0.001005  -5.054 4.56e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

```
Residual standard error: 0.3412 on 3230 degrees of freedom
Multiple R-Squared: 0.1323,  Adjusted R-squared: 0.131

Convergence attained in 2 iterations with relative change 0
****Summary of Davies test****

Davies' test for a change in the slope

data:  formula = ResponseVar ~ age * LeagueIdx ,   method = lm
model = gaussian , link = identity
segmented variable = age
'best' at = 41, n.points = 27, p-value = 1
alternative hypothesis: two.sided

****Summary of Liklihood Ratio test****
comparing models for RCLatencySegmented
Likelihood ratio test

Model 1: ResponseVar ~ age * LeagueIdx
Model 2: ResponseVar ~ age + LeagueIdx + U1.age + psi1.age + age:LeagueIdx
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -1109.5
2   7 -1108.7  2 1.5889     0.4518
```

Segmented analysis with median FAL as response

```
**********************
******* FALMedian ********
**********************
****Summary of linear model****

Call:
lm(formula = ResponseVar ~ age * LeagueIdx, data = DATA)

Residuals:
    Min      1Q  Median      3Q     Max
-429.90  -79.62   -4.46   73.99  764.16

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     690.1490    41.3900  16.674  < 2e-16 ***
age               5.1508     1.7702   2.910  0.00364 **
LeagueIdx2     -128.0362    51.2554  -2.498  0.01254 *
LeagueIdx3     -200.7727    48.2610  -4.160 3.26e-05 ***
LeagueIdx4     -267.8826    47.8738  -5.596 2.38e-08 ***
LeagueIdx5     -321.0412    49.0664  -6.543 6.98e-11 ***
LeagueIdx6     -423.5798    54.2080  -7.814 7.42e-15 ***
age:LeagueIdx2    1.7979     2.2141   0.812  0.41685
age:LeagueIdx3    2.2416     2.0828   1.076  0.28190
age:LeagueIdx4    1.6959     2.0717   0.819  0.41309
age:LeagueIdx5    0.7137     2.1478   0.332  0.73970
age:LeagueIdx6    2.6800     2.4367   1.100  0.27147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 126 on 3263 degrees of freedom
Multiple R-squared:  0.4356,Adjusted R-squared:  0.4337
F-statistic: 228.9 on 11 and 3263 DF,  p-value: < 2.2e-16

****Summary of Segmented model****

***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = lm_seg, seg.Z = ~age)

Estimated Break-Point(s):
   Est. St.Err
26.814  1.476

Meaningful coefficients of the linear terms:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)     743.374     44.862  16.570  < 2e-16 ***
age               2.509      1.973   1.271  0.20369
LeagueIdx2     -134.841     51.226  -2.632  0.00852 **
LeagueIdx3     -207.203     48.234  -4.296 1.79e-05 ***
LeagueIdx4     -286.530     48.195  -5.945 3.05e-09 ***
LeagueIdx5     -354.725     50.048  -7.088 1.66e-12 ***
LeagueIdx6     -470.228     55.915  -8.410  < 2e-16 ***
U1.age            7.057      2.214   3.187       NA
age:LeagueIdx2    2.128      2.213   0.961  0.33646
age:LeagueIdx3    2.559      2.082   1.229  0.21924
age:LeagueIdx4    2.642      2.091   1.263  0.20654
age:LeagueIdx5    2.381      2.203   1.081  0.27989
age:LeagueIdx6    4.987      2.530   1.971  0.04881 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 125.8 on 3261 degrees of freedom
Multiple R-Squared: 0.4375,  Adjusted R-squared: 0.4352

Convergence attained in 2 iterations with relative change 0
****Summary of Davies test**** FALMedian

Davies' test for a change in the slope

data:  formula = ResponseVar ~ age * LeagueIdx ,   method = lm
model = gaussian , link = identity
segmented variable = age
'best' at = 27, n.points = 27, p-value = 0.0107
alternative hypothesis: two.sided

****Summary of Liklihood Ratio test****
comparing models for FALMedian
Likelihood ratio test

Model 1: ResponseVar ~ age * LeagueIdx
Model 2: ResponseVar ~ age + LeagueIdx + U1.age + psi1.age + age:LeagueIdx
  #Df LogLik Df  Chisq Pr(>Chisq)
1  13 -20479
2  15 -20473  2 11.143   0.003805 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```