IDENTIFYING EXPERTISE: DATA EXPLORATION IN TETRIS

By

John K. Lindstedt

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COGNITIVE SCIENCE

Approved:

_____

Dr. Wayne D. Gray, Thesis Adviser

_____

Dr. Michael J. Schoelles, Thesis Adviser

_____

Dr. Brett R. Fajen, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

October 2013
(For Graduation December 2013)

Contents

List of Tables

List of Figures

Acknowledgement

Thanks to Michael Kalsher for his advice on some of the statistical models presented in this paper. The work was supported, in part, by grants N000140910402 and N000141310252 to Wayne Gray from the Office of Naval Research, Dr. Ray Perez, Project Officer.

Abstract

The identification of expertise in a complex task is trivial when all data is in after the fact. To achieve a better understanding of the nature of expertise my first goal, in this work, is to identify the behaviors most predictive of different levels of expertise (novice to expert) in the video game Tetris. The second goal is to examine how little data is required to accurately predict levels of expertise. The present study analyzes potential behavioral indicators of expertise in Tetris at three levels (global, local, and immediate) and under four levels of data quantity. Presented are statistical models using the established metrics of behavior to predict overall output performance. Results indicate mild success in predicting performance, offering a starting point for analyzing other complex and dynamic tasks with explicit criteria for successful performance.

1. Introduction

1.1 Introduction

It seems easy to identify which baseball batters are experts. One can look at their outputs: batting average, fouls, or total runs scored. The trouble is, one can only really make assessments on these outputs after the fact, once all the numbers are in, and the point is somewhat moot. But it seems there must be something discriminating about these experts at a more fundamental level, something identifiable in the way they are playing the game that forms the basis for their continued excellent performance.

What are the hallmarks of the exceptional player's expertise? Is it something about the way they grip the bat, or their stance? Is it in their ability to hit a certain kind of pitch over others? Are they slightly faster to respond, or more deliberate with their actions? Is it because they know when to bunt? Moreover, how much of the player's performance do we need to see in order to make an informed assessment of his or her expertise?

These questions lay the groundwork for asking the question: can we identify *elements of expertise*, behaviors made from instant to instant during performance which will allow us to rank a person on a scale ranging from novice to expert by observing just a thin slice of their behavior? I investigate this question using the video game Tetris.

2. Background

2.1 Study of human expertise

The history of the scientific study of human expertise is nearly as long as the history of scientific psychology, with publications dating back to the discovery of the plateau in skill gain in telegraph operators in 1897 (Bryan and Harter), to an overthrowing of that notion in favor of continuous, if subtle, skill gains throughout the acquisition of expertise (Keller, 1958), and ultimately to a reconciliation of the two findings as valid depending on the measurement device (e.g., Robertson & Glines, 1985).

The conflict between the major claims in the literature highlights the importance of behavioral metrics and the available theoretical constructs at the time. Although Bryan and Harter collected some data with millisecond accuracy, their general methodology lacked a few important controls and their main theoretical construct was stated in intuitive terms. Fifty years later, Keller (one of the foremost behaviorists of his day) had much higher standards for experimental design as well as a theoretical framework, behaviorism, that had no room for unobservable hierarchical structures. Just 30 years after Keller, Robertson and Glines had available to them the hierarchical theories of the information processing theorists as well as an understanding of the ways in which adopting different strategies could lead to differences in performance. As a consequence, unlike Keller when they looked, they found abundant evidence for individual differences in plateaus that seemed to reflect differences in strategies available or discoverable by students with different intellectual backgrounds (i.e., primarily engineers versus humanities students).

The longterm goal of this project is to provide a theoretical account of extreme expertise in dynamic tasks; that is, those which require an integration of real-time decision-making with a (figurative) tight loop among cognition, perception, and action. In a broad sense, the purpose of this course of work is to address and understand concerns with skills at a higher level than typical experimental psychology paradigms, not to the level of decision-making in stock brokers or politicians, but

more to those complex tasks which at least have clear criteria for success. Examples of these interactive skills include laproscopic surgery (Keehner et al., 2004), piloting jet aircraft and helicopters (Proctor, Bauer, & Lucario, 2007; Hays, Jacobs, Prince, & Salas, 1992), and detection of enemy submarines hiding in deep waters (Ehret, Gray, & Kirschenbaum, 2000). For the purposes of a focused study drawing from the undergraduate population of an institution such as RPI, surgeons, helicopter pilots, and submarine commanders would be difficult to access. However, there are plenty of people in the general population who have spent thousands of hours acquiring extreme expertise in video games. These people are the subject of the current study and my first attempt at *thin-slicing* the expertise in Tetris is the subject of this paper.

2.2 Why Tetris?

Tetris is a videogame that is both easy to comprehend and difficult to master. The game is simple in that it has relatively simple rules (introduced in the next section) and players make decisions based on a limited set of potential actions (arranging and placing game pieces). However, there is much for a novice player to learn. The game space changes as a result of decisions made by the player. Errors accumulate and one error tends to lead to another error until catastrophic failure occurs (where "catastrophic failure" in Tetris is simply the end of the current game). As the player succeeds, time pressure increases so that decisions have to be made within decreasing time windows. Furthermore, achieving the highest rewards requires performing maneuvers that risk error and reaching levels of the game where time pressure is highest.

To become highly proficient in the task, players must learn to effectively negotiate the error cost and the increasing time pressure by employing cognitive abilities such as: use of working memory, mental rotations, perceptual comparisons, strategic planning, and prediction, as well as the dexterous and rapid execution of chains of motor commands. Mastering Tetris requires the novice to coordinate the effective and efficient use all of these cognitive resources, abilities, and strategies. For these reasons, we see Tetris as an excellent platform for investigating the acquisition of

expertise in a dynamic, real-time task.

In addition, Tetris has been used to document a variety of cognitive phenomena. A short list includes: epistemic versus pragmatic action (Destefano, Lindstedt, & Gray, 2011; Kirsh & Maglio, 1994), gains in cortical mass and BOLD response (Haier, Karama, Leyba, & Jung, 2009), and near and far transfer (Sims & Mayer, 2002).

2.3 The game of Tetris

(For readers already familiar with the game of Tetris, this section is optional review.)

Tetris is a game of increasingly fast-paced, generative puzzle-solving. When playing Tetris, a player manipulates a series of falling shapes, *zoids*, into an arrangement called the *accumulation* at the bottom of the game space. To score points, the player must *clear rows*. This is accomplished by filling at least one row in the accumulation. The immediate result is that points are scored and the row vanishes from the screen (thereby lowing the height of the accumulation). Since not all zoids fit perfectly together, the accumulation gradually rises as rows begin to go unfilled. When the accumulation reaches the top of the game space, the game ends. As the player clears lines, the game-level increases, speeding up the drop-rate of the zoid, and thus the difficulty, but also offering increased score payoffs for successfully cleared lines. Figure 2.1 illustrates the game screen as a player would see it.

Each zoid is one of seven unique shapes, all consisting of four contiguous block segments. Once a zoid is released into the game board, it begins automatically dropping, traversing the game space top to bottom in 12 seconds initially, down to 2 seconds at the highest difficulty level.

Scoring is nonlinear with respect to the number of lines cleared simultaneously. Initially, clearing 1 line awards 40 points, 2 lines awards 100 points, 3 lines awards 400 points, and clearing 4 lines simultaneously awards an extreme 1200 points. Clearing four lines, in one episode, *scores a Tetris*, and is notable because of its high payoff and difficulty. Points awarded for *a Tetris* are also modified multiplicatively by the current difficulty level.

The version of Tetris developed for the current study, written in Flash, incorporates a robust logging system which captures all game events and states as they occur in real time. These events are detailed in the next section.

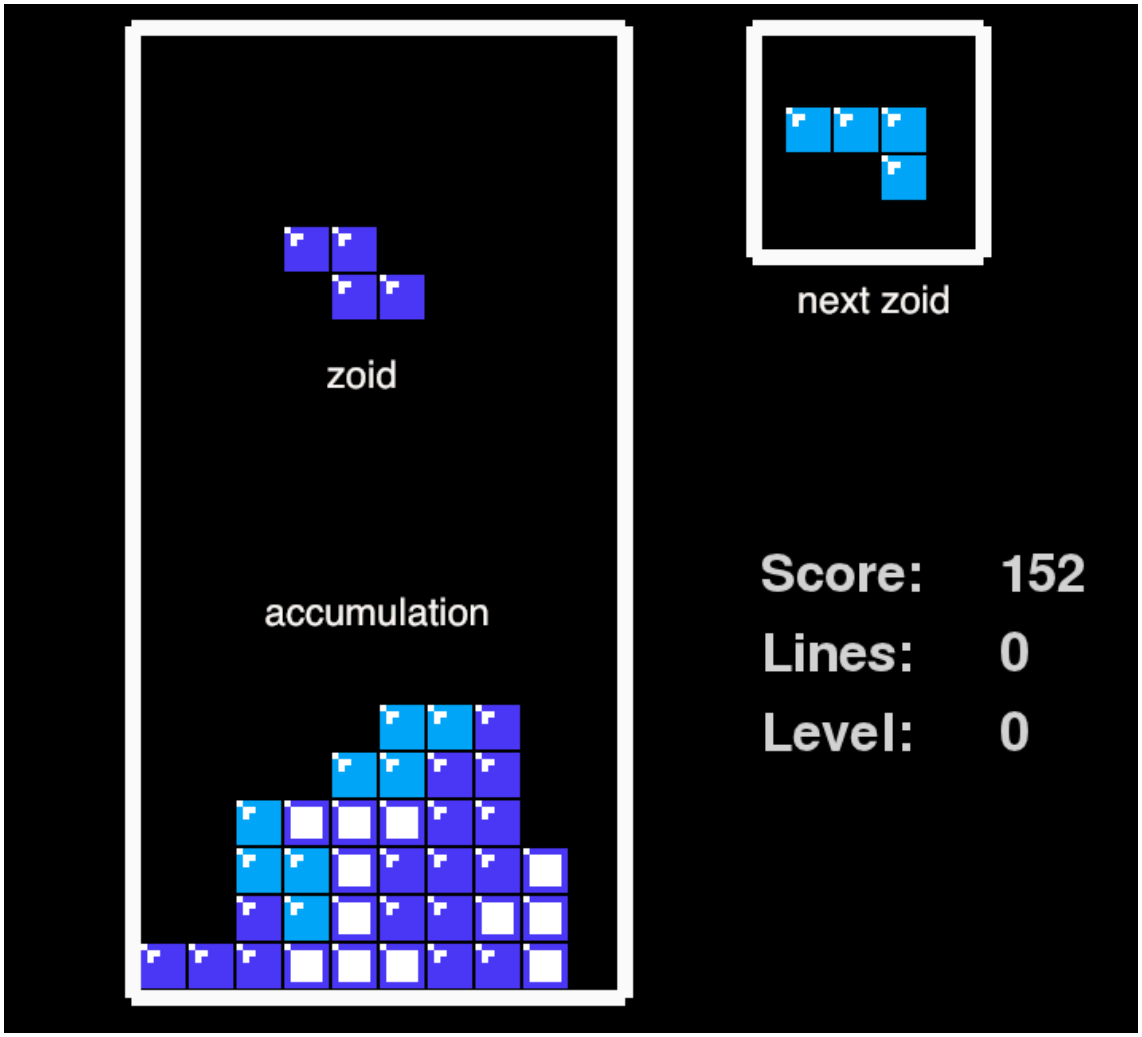

Figure 2.1: Example of the game's task environment. The three relevant game elements are labeled: the current zoid, the next zoid, and the accumulation of previous zoids.

3. Events and metrics

3.1 Events in Tetris

The basic unit of measurement in this study is the *episode*, the time from when a zoid is released until it collides with and locks into the accumulation. It is in this time frame that all measurements of behavior and game state occur.

The player has available three kinds of actions: *rotating* clockwise and counterclockwise, moving a zoid to the left or right (i.e., *translating* between columns), and *dropping* the zoid (increasing the gravity intentionally). System events are any actions performed by the game environment, these include: automatically dropping the zoid due to gravity, clearing filled rows and awarding points, and releasing new zoids. Many of these actions occur within milliseconds of one another, a fact which is captured by time stamping in the continuous logging system.

Although the accumulation changes over time as zoids are placed and lines are cleared, an episode has one unique accumulation with which the player interacts for the duration of the episode. Features of the accumulation are critical for understanding the player's current task status: its *height* determines how close the player is to failure, it may contain unreachable holes or *pits* which, for the game to succeed, must be uncovered (by clearing the rows which cover the pits) and filled, or *overhangs*, which can be thought of as outcroppings that must be filled by moving a zoid into it from its left or right side (a very difficult maneuver, especially for novices).

3.2 Measure of expertise

To assess the behavioral differences of expertise, a quantitative definition is required. Due to the difficulty of achieving high scores in Tetris, and the unlikelihood that a player will score highly simply "by accident," a player's long-term ability to achieve high scores is considered a basic measure of their expertise; that is, a player's expertise is equated to the maximum score the player was able to achieve during any of their games played during data collection. Because scores tend to increase

nonlinearly (later levels award disproportionately more points) and seem to follow a somewhat exponential pattern, the chosen metric of a player's expertise is the base-10 logarithm of their maximum game score.

3.3 Predictive measures

Because the task environment in Tetris is sufficiently simple, many details of task performance can be extracted which may reflect differences in novice and expert behavior. It is important to point out that the target of this study is not only finding those metrics which are the root cause of more expert performance, but also any metrics which reliably co-occur with expert ability. This investigation remains agnostic to this distinction between components and markers of expertise.

These various metrics can be categorized at three successive time scales of human action (Newell, 1990, p. 122): global (10^2 seconds), local (10^1 seconds), or immediate (10^0 seconds). Some examples of the calculated metrics are given in the following subsections. For a list of all metrics calculated for this investigation, consult Appendix A.

3.3.1 Global metrics (10^2 seconds)

These assess the player's overall game status as reflected in the built accumulation. These metrics are associated most closely with *survivability* in the game, such as the overall height of the accumulation, or the number of unworkable holes, or pits, which the player has accrued during play. These metrics, averaged across sections of gameplay, indicate broad patterns of performance which may differentiate between novices and experts, particularly in terms of long-term strategies. Figure 3.1 illustrates three of these metrics.

Average height: The average of all column heights in the accumulation, indicating how close the player is to failure.

Pits: The total number of unworkable pits (covered empty spaces) present in the accumulation that must be uncovered for the player to reduce its height. These equate with

Levelness: Measures the relative flatness (or jaggedness) of the top of the top contour of the accumulation by summing the changes in height from column to column.

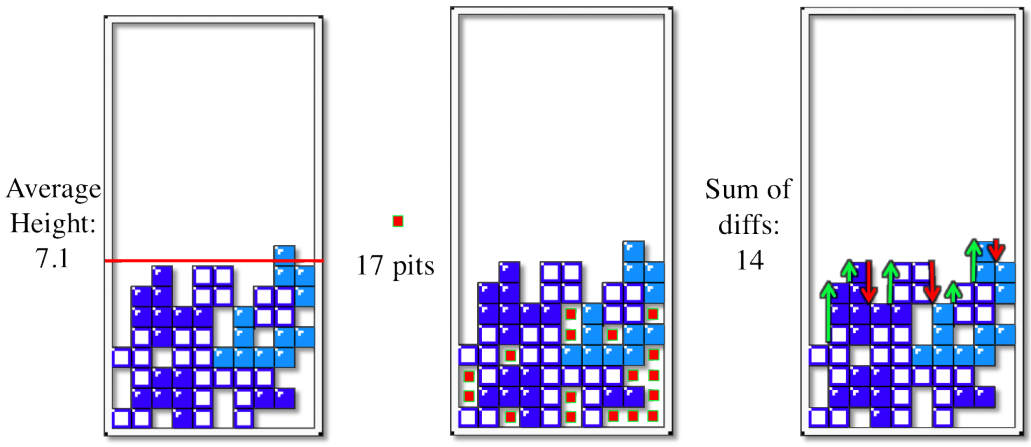

Figure 3.1: A visual depiction of the methods used for characterizing a given accumulation's average height (left), pits (center), and levelness (right).

3.3.2 Local metrics (10^1 seconds)

These assess the kinds of zoid-placements the player selects in relation to possible positions on the accumulation. This includes features such as the number of perimeter segments matched during a placement (i.e., does that zoid fit flush with its surroundings, or does it stick out precariously?), or whether the placement creates pits or overhanging segments in the accumulation which complicate later gameplay decisions. Zoid placements are also compared across all potential placement locations and orientations for the current zoid, giving a ratio of assumed "goodness" for a placement. These local metrics account for the kinds of decisions made at each step of the game. Figure 3.2 illustrates the calculation of two of these metrics.

Matched edges: The number of segments of the placed zoid which are touching the surrounding accumulation or walls.

New pits: The number of new pits created by this move, a measure of error commission.

Uncovered pits: The number of pits uncovered by this move, a measure of error recovery.

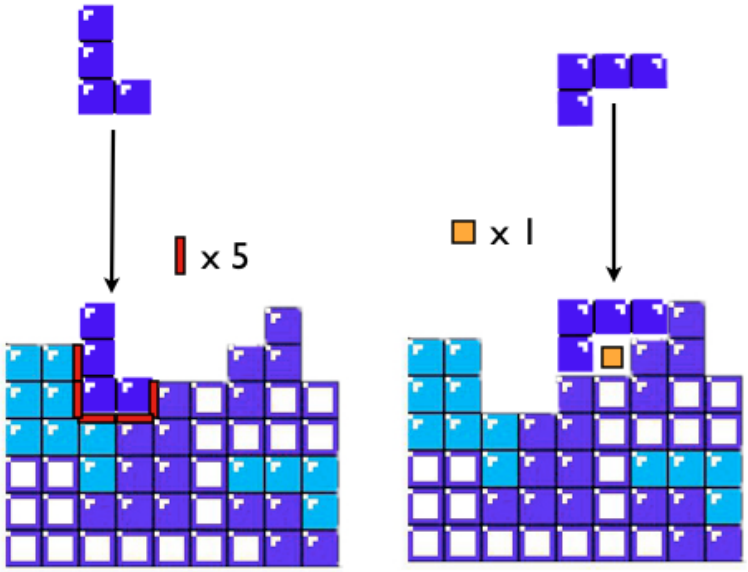

Figure 3.2: A visual depiction of the methods used for characterizing a given zoid placement's matched edges (left) and creation of new pits (right).

3.3.3 Immediate interaction metrics (10^0 seconds)

These account for how a zoid placement is executed, what can be thought of as the sensory-motor aspects of the gameplay. These include measurements of reaction times for various actions, such as the first keypress in an episode, and the first commission of a zoid drop to indicate that a decision has been made. These measures account for the rapid interactive skills a player employs to perform the basic decisions in the local metrics.

Grouped actions: The number of clusters of similar actions performed in sequence (i.e., 3-translations, 2-rotations, 16-drops). This measure reduces the sequences of actions to more conceptually coherent segments, with lower numbers implying less scattered executions.

Average latency: The average time between actions taken by the player.

Drop latency: The time from the start of the episode until the player decides to drop the zoid.

Each of these metrics is tallied and recorded once per episode. By examining elements from these three categories of performance, a broad, detailed picture of each player's gameplay is captured as it occurs in real time.

3.4 Task dependent vs. task independent measures

In understanding the expertise of those participating in a complex task, it bears acknowledging that there exist individual differences among players that would contribute to many aspects of their skilled performance. Such differences could potentially be captured by task-independent measures of general faculties of participants from a battery of cognitive (i.e. working memory span) or physiological (i.e. manual dexterity) tasks. Although these task independent measures would be of great interest to the general pursuit of understanding expertise, this investigation maintains focus only on task-dependent measures, in this case those which are relevant specifically to the game of Tetris. This is initially a pragmatic concern: the manner in which data were collected for this study afforded little opportunity for administering a cognitive task battery. Still, beyond the issue of implementation, the question remains as to whether there are observable differences in task dependent behavior alone, in spite of any such individual differences.

4. Methods

4.1 Data collection

To acquire data from a cross section of players with different levels of expertise, the CogWorks laboratory sponsored a Tetris tournament at Rensselaer Polytechnic Institute's *Genericon* – a convention for gaming, comics, Japanese anime, and all things "nerd culture." Participants in the tournament were volunteers from the pool of all those attending the convention, comprised primarily of RPI undergraduates.

Before the tournament, participants played two rounds of Tetris to determine their eligibility for competing. Once entered, participants competed in pairwise elimination matches wherein the highest score wins. The top three players of each tournament were offered a cash prize, provided they submitted to participate in a laboratory examination wherein they played an additional hour of Tetris.

Data were collected data from all tournament players using this procedure at two successive *Genericon* events in 2006 and 2007.

4.2 Data reduction

At the end of data collection, the body of raw data consisted of 178 full games of Tetris played by 64 unique players, with game scores spanning five orders of magnitude (less than 100 points to over 1,000,000; see Figure 4.1). The raw data captured during these games contained over 2.3 million lines of state and event log data. As a result of calculating the predictive measures described in the previous section, 45,237 behavioral observations in the form of episodes remained.

4.3 Data filtering

Games wherein a player did not clear any lines were omitted from analysis, as these represent sessions which were either aborted or wherein the player clearly did not understand the game rules. Additionally, players were sometimes observed self-aborting games by rapidly dropping zoids until a game-over was achieved. These

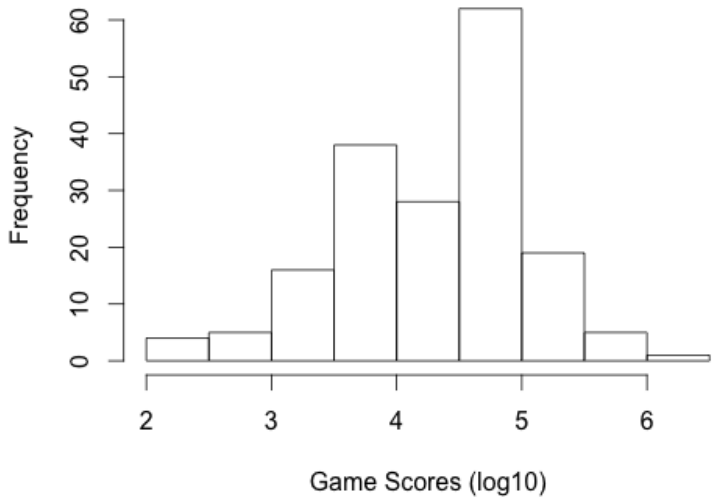

Figure 4.1: Histogram of all game scores from all players from both Genericon Tetris Tournaments held in 2006 and 2007.

episodes were omitted from analysis, as they reflected gameplay behavior with misaligned goals. After filtering, data from 57 of the 64 players remained.

4.4 Observation window

An important consideration for the present data set is that it is naturalistic: no experimental controls were put in place, and no manipulations were made to the basic game. As such, there is a great deal of unevenness in the data set. The task environment is influenced greatly by the randomness of the zoid selection and player strategy, as is the number of episodes it takes a player to advance to the next difficulty level (where game speed is increased), or even the number of episodes played before the game ends. To control these elements would be to interfere with the basic structure of the game and deviate from the way players would naturally approach the game, hindering our ability to find natural expert players "in the wild" as such. Thus, for this investigation these vital game elements are left uncontrolled, and instead a *moving window* is used to examine the gameplay data.

A key element of this exploration is whether one can *thin-slice* by predicting expertise from a relatively small amount of data. Across all subjects and games, the mean number of episodes per game was 264.74 [Median = 237, Min. = 41,

Max. = 1388, S.D. = 210.97]. For the purposes of thin-slicing, in all cases the observation window begins with the 1st episode of each game, wherein all players have a completely empty accumulation with which to work. For each player, data are averaged for all games for episodes 1-2 (an extremely thin slice of behavior), 1-10, 1-100, and all (using all available data for the analysis), illustrated in Figure 4.2. Averaging behavioral measures across this window results in aggregate measures of performance which are representative of a player's behavior for the chosen observation window (Figure 4.3). The present question is whether measures made on these different slices of performance are predictive of overall performance.

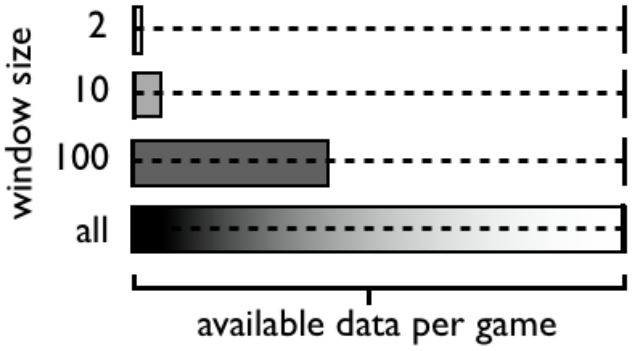

Figure 4.2: Illustration of the chosen observation window sizes relative to the available data per game.

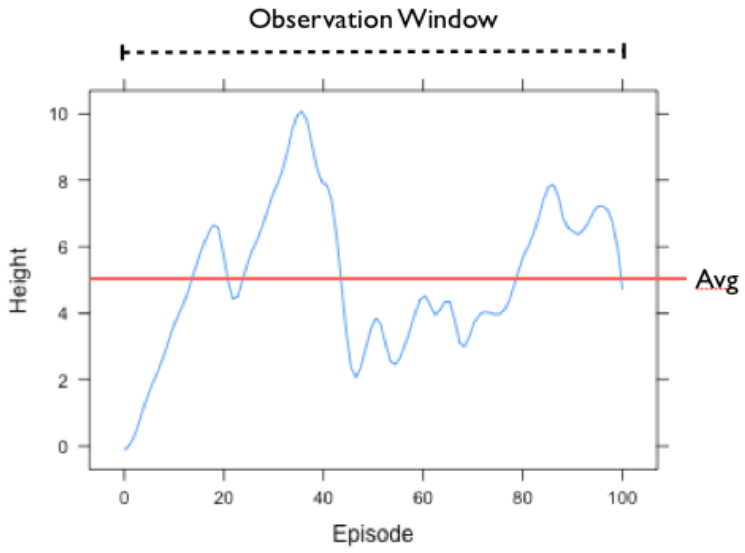

Figure 4.3: Example of the characterization of behavior scores over the observation window (100 episodes shown) via averaging.

5. Results

5.1 Multiple linear regression models

Prior to modeling, the dataset was sampled using a simple random assignment, using 80% of the data for training and leaving 20% for testing model predictions. The samples were verified as having similar distributions for the dependent measure of expertise [Training set: Mean = 4.43, S.D. = 0.61; Test set: Mean = 4.51, S.D. = 0.73].

For each of the four selected observation window sizes (2, 10, 100, and all episodes), a backward stepwise multiple regression was conducted on each training data set using all predictors detailed in the Predictive Measures section. Table 5.1 shows the results of these models, and Table 5.2 illuminates the significance of each model's predictors. Note that the number of predictors ultimately used in each model varies due to the stepwise selection process. Figure 5.1 shows the fit of each model to the training data.

5.2 Prediction

To assess each model's ability to predict unseen data, predictions were performed on the test data set (20 percent of observations). The Predictions section of Table 1 shows the relative success of each model as determined by the fit of a Pearson's product-moment correlation. Figure 5.2 shows the fit of the test set data to the model predictions.

This chapter previously appeared as: Lindstedt, J. K. and Gray, W. D. (2013). Extreme expertise: exploring expert behavior in Tetris. In Knauff, M., Pauen, M., Sebanz, N., and Wachsmuth, I., (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Table 5.1: Results of linear regression model for all window sizes.

	Observation window size			
	2 eps	10 eps	100 eps	all eps
Multiple R^2	.4607	.3913	.5882	.8185
Adjusted R^2	.3686	.2509	.5058	.7767
DF	(7,41)	(9,39)	(8,40)	(9,39)
F-value	5.003	2.786	7.141	19.55
p-value	<0.001	0.01	<0.0001	<0.0001
Prediction				
Correlation	0.344	-0.235	0.697	0.757
p-value	0.27	0.46	<0.02	<0.01

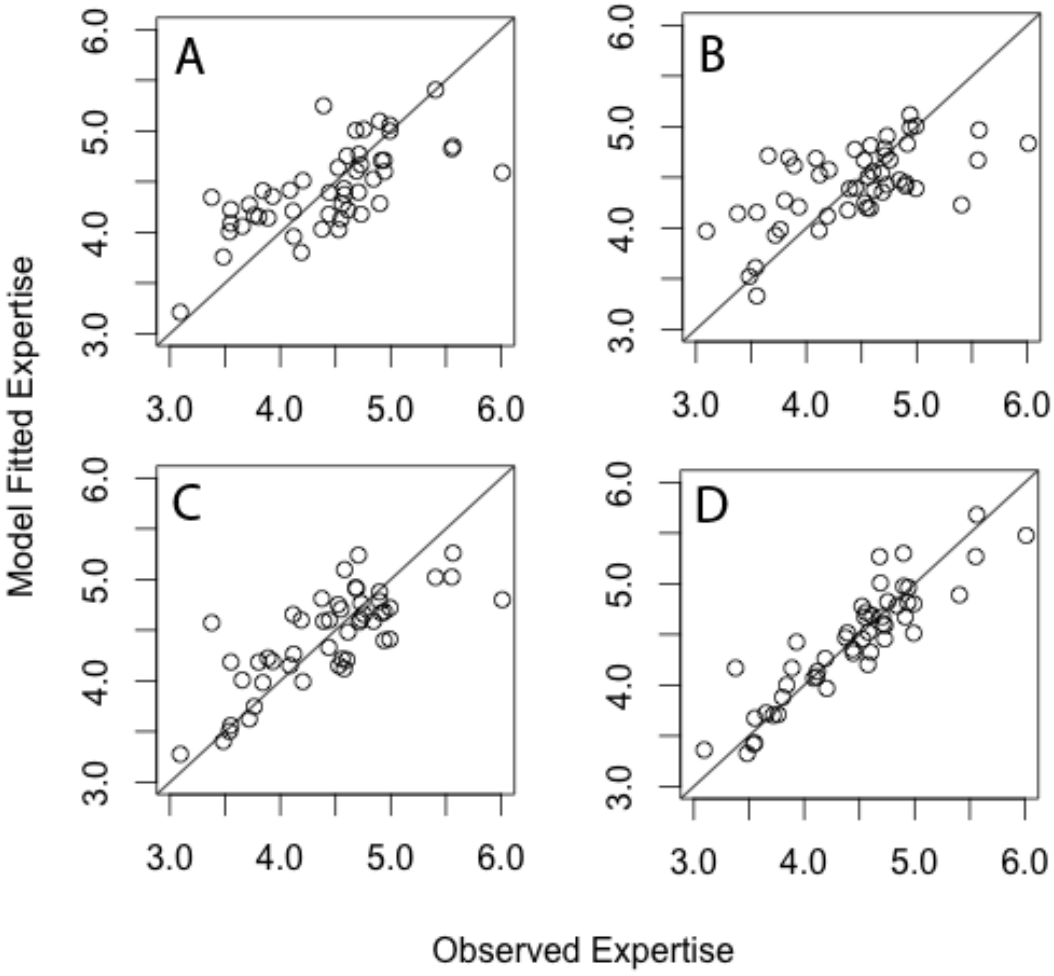

Figure 5.1: Plot of observed expertise values in training data against the fitted values from the multiple regression models. Different plots for models sampling from A) 2 episodes, B) 10 episodes, C) 100 episodes, and D) all observed episodes per game.

Table 5.2: List of significant predictors across models of differing observation window sizes. Significance codes are: '*' - p <0.05; '' p <0.01; '***' p <0.001; '.' = present but not significant.**

	Window Size (episodes)			
	2	10	100	All
Intercept	.	.	**	.
Global metrics:				
Average Height		*		
Pits	*		.	*
Overhangs		.		
Roughness				
Levelness	*			
Spire	**			
Tetris progress				
Zoid-positions	*			
Local metrics:				
Matched edges		*	**	*
Match ratio	***			
New pits		*		
Uncovered pits		*	.	**
Filled overhangs	**		.	***
Current zoid-positions		.		
Immediate metrics:				
Total translations			.	.
Total rotations				*
Grouped actions	*			
Drop ratio		.		***
Initial latency		.		.
Average latency			*	
Drop latency		*	**	***

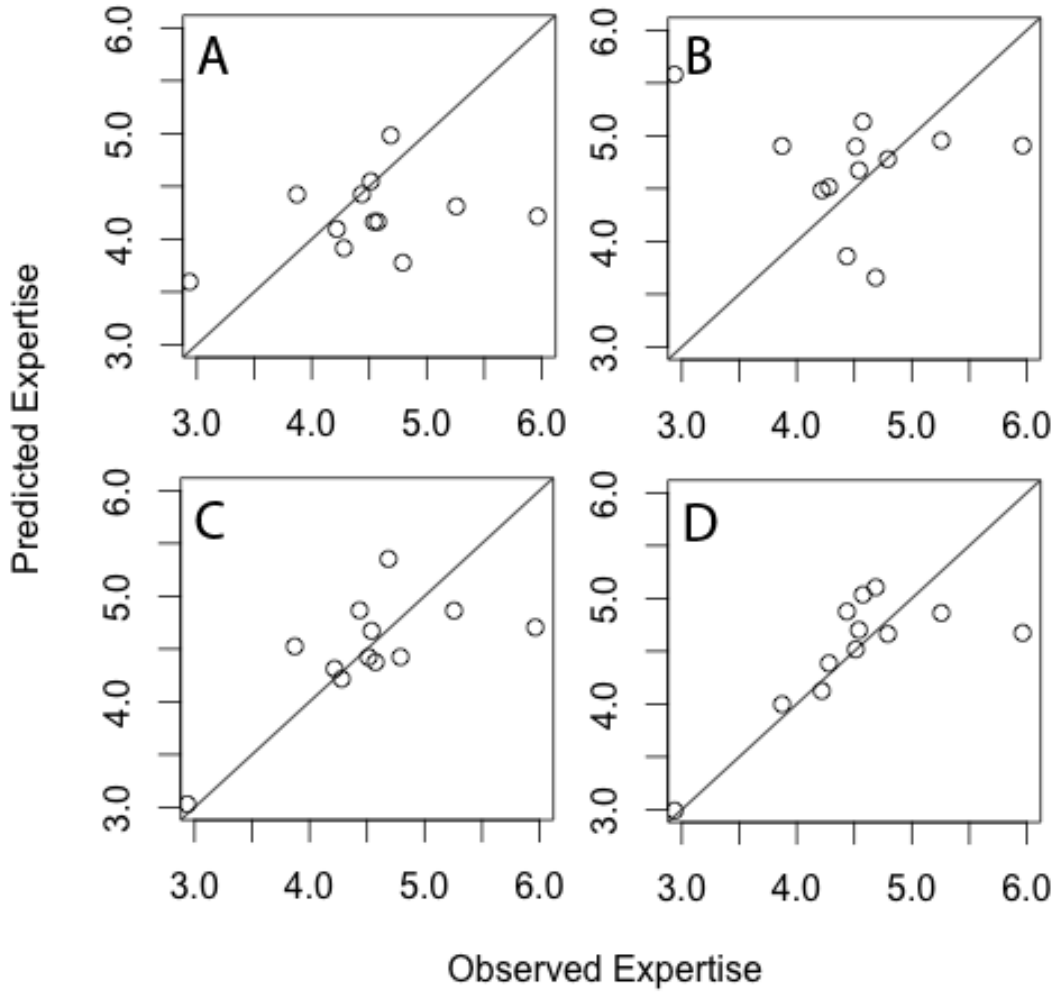

Figure 5.2: Plot of observed expertise values in test data set against the values predicted by the multiple regression models. Different plots for models sampling from (A) 2 episodes, (B) 10 episodes, (C) 100 episodes, and (D) all observed episodes per game.

6. Discussion

These results show significant fits for models created using all sizes of observation windows, from data spanning just two episodes to the use of the entire data set. The two models sampling from just 2 and 10 episodes each are notable for their good fits, but both ultimately fail to predict unseen data. Nonetheless, their fits are encouraging in that they achieve a measure of success even when based on such a small proportion of the player's observable performance data.

Models sampling from more data are naturally able to account for more of the variance in the data, as seen by the increasing adjusted R^2 values for those models with larger windows, with the model sampling all data presumably demonstrating a maximum of success. Interestingly, the model sampling only the first 100 episodes (less than a quarter of all observed data) maintains a fit to the training data and ability to predict the test data comparable to that of the model sampling all data. This, too, is encouraging in the pursuit of using small proportions of data to predict long-term performance.

Encouraging as the relative successes of these models may be, their contribution to the theory of expertise is limited. As previously mentioned, the set of potentially predictive features included in the models are fully agnostic to causation; there is no distinction between components of expertise and markers of expertise. As such, we cannot effectively hypothesize whether, say, an expert player is more agile because they are better at the game (a marker of expertise), or they are a better player because they began or became more agile (a component of expertise).

6.1 Feature co-dependence

It is tempting to draw conclusions from the lists of significant predictors presented in Table 5.2, but there is, regrettably, a non-trivial sampling effect; depending on how the data set is partitioned into training and test sets, these significant variables tend to shift, vanish, and reappear on subsequent samplings. This is likely due to two underlying effects: a strong effect of individual differences, as suggested by

Robertson and Glines (1985); and a high level of correlation between these variables, because many of them necessarily depend on one another (e.g., average height being necessary for Tetris progress). These results do not yet account for these covert effects and so no strong conclusions may be drawn about the individual predictors' viability in predicting long-term Tetris performance. I do, however, offer two points of speculative commentary based on observation of these effects: first, some predictors seem to emerge as significant more frequently than others, and second, predictors representing all three time scales (global, local, and immediate) tend to emerge as significant across samplings, indicating that there may be something unique about each of these categories which contributes independently to performance.

To address the issues of feature co-dependence, future work with the present data set will involve, first, expanding the set of behavioral predictors beyond only averaging across the window using variability (i.e. standard deviation of a predictor) and time series analysis (i.e. examining the pattern of a predictor over the course of the game). Second, more sophisticated and automated feature selection algorithms will be employed to determine which of the expanded set of behavioral predictors underlie expert performance. I expect this two-step process will produce a more robust picture of behavior in the game of Tetris, and thus a better lens through which to examine performance.

6.2 Additional applications

What can be taken away from the results of this investigation is that a careful analysis of a complex, dynamic task's structure can lead to an understanding of how experts perform compared to novices in the task, and, importantly, such differences can be captured with sparse performance data. Beyond Tetris, this approach can be used to examine any task of similar complexity, provided said task has clear criteria for successful performance, and said performance produces continuous and measurable outputs.

An example of a similar task to which this paradigm can be applied is the *Space Fortress* task. *Space Fortress* is a fast-paced game in which the player steers a ship through a frictionless space and attempts to score points by shooting a limited

supply of missiles at a central target while avoiding hazards on the screen. The game has a number of distinct ways in which the player can increase points: flagging and destroying mines, destroying the central target, collecting bonuses, among others. Following a similar approach to the present study, one could classify a number of possible behaviors in Space Fortress that contribute or detract from successful performance (i.e. hits, errors, recoveries) across the global, local, and immediate time scales laid out in this paper (i.e. missing targets, taking erratic paths, or running out of ammunition). In doing so, one would produce a robust set of behavioral predictors which could be used to distinguish the task-dependent features most important to expert performance in Space Fortress.

6.3 Future studies

Two subsequent lines of research follow directly from the present study: a series of focused experimental studies and a computational cognitive model. The experimental studies will seek to address the causative relationship between the behavioral predictors discussed here and expertise. Some of the present findings suggest that certain aspects of gameplay go hand-in-hand with expert performance, for instance, measures pertaining to manual dexterity in the task seem important in predicting performance. The experimental studies will use a modifiable version of Tetris to investigate the effects of removing these aspects of the game, in this case the time-pressure of the falling blocks, to examine both the effects of such manipulations on existing experts and novices, but also their effects on the rate of expertise acquisition.

In parallel, I will work to construct a computational cognitive model of the Tetris task in the ACT-R architecture. After creating a model that can achieve some measure of baseline performance in the game, it will split into two separate models: a novice and an expert. The key distinction between the two will be in how each model uses the information in the world (i.e. the available features from the *global* and *local* levels), serving as a theoretical test of some of the distinguishing characteristics of novices and experts presented here. The long-term goal of this line of research would be to produce a model that is capable of transversing from novice

to expert levels of play in a way that mirrors human skill acquisition.

7. Conclusions

The goal of the present work is to identify the elements of expertise that predict the continuum of performance in the game of Tetris. As a first step, I collected data from a wide variety of Tetris Tournament players and used it to derive metrics of global, local, and immediate interactions. Here I reported my first statistical models of these data and their initial success at predicting level of expertise from thin-slices of behavior.

Although the results are tentative, these initial successes are promising in applying a general cognitive task approach to extreme expertise. The presented categories of global, local, and immediate interaction are based on three successive levels of the *time scale of human action* (Newell, 1990). At least some of the initial items for each scale shows some success as a predictor of expertise. Thin-slicing seems to produce valid predictions as even the regression model based on the first two episodes of each game had some predictive validity. In light of these initial successes the project will continue with the collection of an order of magnitude more data from an order of magnitude more players at all levels of expertise.

The predictive modeling used in this paper has thus far been limited to the statistical technique of multiple regression. I am actively investigating more robust classification and prediction techniques. Further work will also seek to address the individual differences across players at the same skill level and will attempt to extract a more refined set of metrics of behavior with fewer co-dependencies through the use of more refined feature selection techniques. Additional experimental studies are planned to target and alter specific behavioral features of the game to see how the alterations both affect existing expertise and encourage the acquisition of expertise.

Bibliography

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

Destefano, M., Lindstedt, J. K., & Gray, W. D. (2011). Use of complementary actions decreases with expertise. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2709–2014). Austin, TX: Cognitive Science Society.

Ehret, B. D., Gray, W. D., & Kirschenbaum, S. S. (2000). Contending with complexity: developing and using a scaled world in applied cognitive research. *Human Factors*, *42*(1), 8–23.

Haier, R., Karama, S., Leyba, L., & Jung, R. (2009). Mri assessment of cortical thickness and functional activity changes in adolescent girls following three months of practice on a visual-spatial task. *BMC Research Notes*, *2*(1), 1–7. Retrieved from http://dx.doi.org/10.1186/1756-0500-2-174

Hays, R. T., Jacobs, J. W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness: a meta-analysis. *Military Psychology*, *4*(2), 63–74. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/s15327876mp0402_1

Keehner, M., Tendick, F., Meng, M., Anwar, H., Hegarty, M., Stoller, M., & Duh, Q. (2004). Spatial ability, experience, and skill in laparoscopic surgery. *American Journal of Surgery*, *188*(1), 71–75.

Keller, F. S. (1958). The phantom plateau. *Journal of the Experimental Analysis of Behavior*, *1*(1), 1–13.

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513–549.

Lindstedt, J. K. & Gray, W. D. (2013). Extreme expertise: exploring expert behavior in Tetris. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Proctor, M. D., Bauer, M., & Lucario, T. (2007). Helicopter flight training through serious aviation gaming. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, *4*(3), 277–294. Retrieved from http://dms.sagepub.com/content/4/3/277.abstract

Robertson, R. J. & Glines, L. A. (1985). The phantom plateau returns. *Perceptual and Motor Skills*, *61*(1), 55–64.

Sims, V. K. & Mayer, R. E. (2002). Domain specificity of spatial expertise: the case of video game players. *Applied Cognitive Psychology*, *16*(1), 97–115.

Appendix A
Performance metrics

All metrics are measured continuously, at least once per zoid episode. These are then averaged across the observation window and across all games played by a single player, resulting in a single characteristic score per metric per player.

A.1 Global metrics (10^2 seconds)

Global metrics are concerned with the state of the accumulation in the game. These are metrics that take into account the longest time frame of behavior, and are rather stable across episodes; that is, fluctuations tend to come slowly.

A.1.1 Average height

The average height is calculated by summing all column heights and dividing by the width of the game board. High values are associated with the accumulation approaching the top of the game space (and impending game over), while lower values are "safer." This metric is insensitive to the actual shape of the accumulation.

A.1.2 Pits

A pit is any empty space in the board that is closed off in all directions, i.e. there is no open path to the top of the board. Pits are troublesome to players, as lines they occupy cannoy be cleared until the zoid segments covering them have been cleared. An excess of pits contributes heavily to a game over.

A.1.3 Overhangs

An overhang is any empty space that *does* have an open path to the top of the board. Like pits, these formations are still difficult to deal with, but a skilled player can maneuver a zoid to fill them in.

A.1.4 Roughness

Roughness is a measurement of the "randomness" of the accumulation. It is calculated by summing the number of contiguously filled or unfilled segments in each row or column, and summing those results. An accumulation with no roughness is empty (or completely filled!), whereas an accumulation with a high roughness score may have many pits, overhangs, or be very uneven. More orderly, minimally rough boards are easier to deal with.

A.1.5 Levelness

Levelness is measured by summing the differences between column heights. This gives an output that is very high for jagged accumulations, and very low for accumulations with a flattened top. A perfectly flat accumulation is actually not entirley ideal, but an excessively jagged one is also unworkable; some amount of unevenness affords more opportunities for good zoid placements.

A.1.6 Spire

The spire of an accumulation is a simple measure of how tall the largest "spike" structure is compared to the average height of the accumulation. This measure is low for flat accumulations and high for accumulations featuring a large spike, i.e. the player has piled many zoids in one or two columns.

A.1.7 Tetris progress

Tetris progress is measured by the number of contiguous rows with the same 9 columns filled, i.e. where an I-shaped zoid could be placed to clear multiple lines at once. A tetris progress measure of 4 is the minimum required to be able to clear 4 lines at once, granting the most points possible. As this value increases, the player is building the accumulation higher and higher (and more dangerously) in preparation for the large payouts offered by the occurrence of an I-zoid.

A.1.8 Any-zoid placements

This metric assesses how many "good," that is non-pit-creating zoid placements, can be made by any zoid for the current accumulation. This is a rough

metric of how useful the accumulation is given all possible zoids. A high value indicates a large number of "good" positions are available and, hence, the accumulation is useful.

A.2 Local metrics (10^1 seconds)

Local metrics are concerned with the features relevant to the placement of a zoid. These features are presumably what drive the player's decisions about where to ultimately place each zoid.

A.2.1 Matched edges

The number of matched contours is the number of side-segments of the current zoid in contact with a wall or accumulation segment when the zoid is ultimately placed. Low values indicate the zoid was placed precariously, whereas high values indicate the zoid has been placed in a well-fitting position. [figure needed]

A.2.2 Match proportion

The match proportion compares the current placement's matched edges score against all other possible placements for the current zoid. Values approaching 1 indicate the optimal choice was made, whereas values approaching 0 indicate there was a better option among the available positions.

A.2.3 New pits and uncovered pits

New pits is a volatile, instantaneous measure of whether the current zoid placement has created a new pit. This differs from the global measure of pits in that it is sensitive only to new pit occurrences, and does not continously track all preexisting pits. A high value indicates the player more often creates pits with their placements, while a low value indicates the opposite. Uncovered pits measures the number of pits that have been re-opened for access this episode. An uncovered pit is equivalent to error-recovery. High values indicate more common occurrence. Taken together, new pits and uncovered pits account for the global pits score, but

individually focus on two different and relevant dimensions: error commission and error recovery, as opposed to the simpler errors accrued offered by the global metric.

A.2.4 Filled overhangs

Filled overhangs measures the number of zoid-segments of the current zoid placement which are filling in an overhang. This is a mild form of error recovery, similar to uncovered pits, but also requires a certain amount of dexterity not present in all players.

A.2.5 Current-zoid placements

Like the any zoid placements metric, this measure counts the possible number of non-pit-producing zoid placements, but for the current zoid only. A high value here actually indicates more "good" moves for the current zoid. Interestingly, an important implication of this metric is that it is determined by the *previous* zoid placement, where the player was aware of what the next zoid was (the now-current zoid). Consistently higher values in this metric show that a player is making more accommodations for the next zoid in their placement decisions.

A.3 Immediate interaction metrics (10^0 seconds)

Immediate metrics are concerned with how the player gets the zoid from the top of the game screen to its ultimate destination. This includes measures of dexterity, speed, and path planning and execution.

A.3.1 Total translations

This is the total number of translations performed in an episode. A high number indicates the player either moved the zoid a far distance, indecisively changed strategies, or performed an epistemic sort of action using extra translations to verify the zoid's position.

A.3.2 Total rotations

Much like total translations, this is the total number of rotations performed in an episode. Higher values indicate more rotations. Interestingly, no zoid ever

needs to be rotated more than 3 times to achieve its ultimate orientation, so values beyond this are either in error, indecision, or indicate the use of epistemic actions.

A.3.3 Grouped actions

This measure groups actions into related strings of simple actions (i.e. 3-translations then 1-rotation then 12 drops), and then counts the number of those groups. This is useful in pulling meaning from the raw actions, as many actions can result from a single key-press. High values indicate the player performed many more actions, likely exhibiting strategy change or error-making behavior, while low values indicate a sort of "no-nonsense" approach to placing the zoid.

A.3.4 Drop ratio

A zoid moves discretely through the game board, one row at a time. A player can initiate a "drop" to speed this along once they have decided on their destination. This metric compares the number of player-controlled drops to the total number including those initiated by the game's gravity pulling the zoid downward. A value approaching 1 indicates all of the downward motion of the zoid this episode was intended by the player. A value approaching 0 indicates the player never initiated a drop, either due to hesitation or desire to use all possible time for planning.

A.3.5 Initial latency

This is the time in milliseconds between the start of the episode and the first keypress by the player. Low values indicate the player took less time to process the situation before acting.

A.3.6 Average latency

This is the average time between actions the player takes during an episode. A player with a small value for average latency demonstrates the ability to execute chains of actions quickly.

A.3.7 Drop latency

This is the time it takes the player to decide on a position and maneuver the piece before initiating a "drop." Unlike drop ratio, drop latency is insensitive to stuttered dropping and indecisiveness, but gives a cleaner measure of "time to decide" in a given episode.