Extreme Expertise: Exploring Expert Behavior in Tetris

John K. Lindstedt (lindsj3@rpi.edu) & Wayne D. Gray (grayw@rpi.edu)
Cognitive Science Department
Rensselaer Polytechnic Institute

Abstract

Expertise is easy to identify in retrospect. It is the most expert player who wins the meet and the most proficient team that wins the playoffs. However, sometimes during play we see a masterful move that clearly separates one player from the competition. Our goal, in this work, is to identify the masterful moves or *elements of expertise* that predict the continuum of performance in the game of Tetris. As a first step we have collected data from a wide variety of Tetris Tournament players and used it to derive metrics of global, local, and immediate interactions. Here we present statistical models of these data and report the initial success of these models at predicting level of expertise.

Keywords: expertise, skill acquisition, exploratory analysis, videogames, regression modeling, thin-slicing

Introduction

It seems easy to identify which baseball players are experts. We can look at their outputs: batting average, fouls, or total runs scored. The trouble is, we can only really make assessments on these outputs after the fact, once all the numbers are in, and the point is somewhat moot. But there must be something different 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? We investigate this question using the video game Tetris.

Background

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).

Our reading of the historical literature is that the discrepancy of major claims about the nature of expertise highlights the importance of metrics and of the available theoretical constructs. 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).

Our longterm goal 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. Examples of such 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). Of course, we lack access to surgeons, helicopter pilots, and submarine commanders. However, we do have people who have spent thousands of hours acquiring extreme expertise in videogames. These people are the subject of our study and our first attempt at *thin-slicing* the expertise in Tetris is the subject of this paper.

Why Tetris?

Tetris is a videogame that is both easy to comprehend and difficult to achieve mastery over. 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 (i.e. the end of the game) occurs. 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).

The Game of Tetris

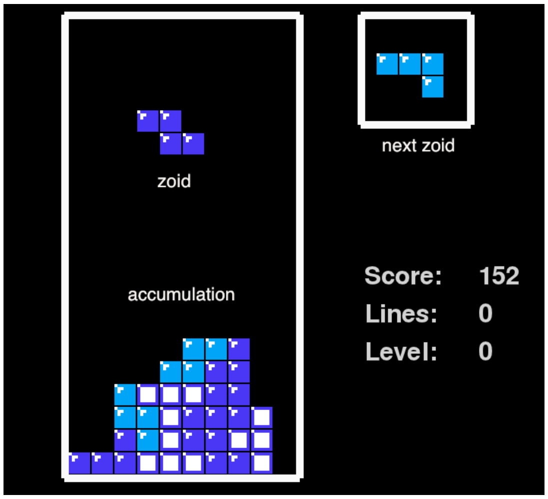
(For readers already familiar with the game of Tetris, this section is optional review.)

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 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 simultaneously *scores a Tetris*, and is notable because of both its high payoff and difficulty. Points awarded for *a Tetris* are also modified multiplicatively by the current difficulty level.

Our version of Tetris, 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.



Score: 152
Lines: 0
Level: 0

Figure 1. Example of the game environment.

Events and Metrics

Events in Tetris

Our basic unit of measurement 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 our continuous logging system.

Though the accumulation changes over time as zoids are placed and lines are cleared, during an episode the player interacts with one unique accumulation. 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's continued success, must be uncovered (by clearing the rows which cover the pits) and filled, or *overhangs* (which can be thought of as a little cave that must somehow be filled by moving a zoid into it from its left or right side *a very difficult maneuver*, especially for novices).

Measure of Expertise

To assess the behavioral differences of expertise, we must define it quantitatively. Due to the difficulty of achieving high scores in Tetris, and the unlikelihood that a player will score highly simply "by accident," we consider a player's long-term ability to achieve high scores 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, our metric of a player's expertise is the base-10 logarithm of their maximum game score.

Predictive Measures

Because the task environment in Tetris is sufficiently simple, we are able to extract many details of task performance which may reflect differences in novice and expert behavior. It is important to point out that we are not searching only for 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.

Our various metrics can be categorized at three successive time scales of human action (Newell, 1990, p. 122): global (10^2), local (10^1), or immediate (10^0).

Global metrics. 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.

Average height: The average of all column heights in the accumulation.

Pits: The total number of unworkable pits (covered empty spaces) present in the accumulation.

Overhangs: The number of covered spaces into which a player may still dextrously maneuver a zoid.

Roughness: A measure of the "randomness" of the accumulation.

Levelness: Measures the relative flatness of the top of the board.

Spire: The difference between the highest point in the accumulation and the average height.

Tetris progress: The number of nearly-filled rows presently lined up in the accumulation, ready to produce high-scoring line-clears.

Zoid-positions: The amount of "good" positions available for any kind of zoid. This is a rough measure of the functional "goodness" of the accumulation the player has built.

Local metrics. 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.

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

Match ratio: Ratio of the number of matched edges to the maximum possible for all positions the zoid could have been placed this episode.

New pits: The number of new pits created by this move.

Uncovered pits: The number of pits uncovered by this move.

Filled overhangs: The number of overhang cavities filled by the current move.

Current zoid-positions: The number of "good" positions available for the current zoid, which may indicate a player's planning for the next zoid in the previous episode.

Immediate interaction metrics. 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.

Total translations: The number of times a zoid was moved left or right in the episode.

Total rotations: The number of rotation actions performed on the zoid this episode.

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.

Drop ratio: The proportion of the zoid's downward movements (in this episode) that can be attributed to the player's intentional dropping versus the system's automatic dropping.

Initial latency: The time (in milliseconds) between the start of the episode and the first action taken by the player.

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, we hope to capture a broad, detailed picture of each player's gameplay as it occurs in real time.

Methods

Data collection

To acquire data from a cross section of players with different levels of expertise, we 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 came to the laboratory and played an additional hour of Tetris.

We collected data using this procedure at two successive Genericons in 2006 and 2007. At the end of data collection, we had data from 57 unique players, with game scores spanning six orders of magnitude (less than 100 points to over 1,000,000).

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, we sometimes observed players self-aborting games by rapidly dropping zoids until a game-over was achieved. These episodes were omitted from analysis, as they reflected gameplay behavior with maligned goals.

Observation window

An important consideration for our 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, we leave these vital game elements uncontrolled, and instead institute a *moving window* through which to examine the gameplay data.

A key element of this exploration is whether we 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 [Min. = 41, Max. = 1388, S.D. = 210.97]. For our 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, we then averaged the data 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). 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. *Our question is whether measures made on these different slices of performance are predictive of overall performance.*

Results

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), we conducted a multiple regression on each training data set using all predictors detailed in the Predictive Measures section. To account for any suppressor effects, a backward step-wise selection process was used in the regression. Table 1 shows the results of these models, and Table 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 2 shows the fit of each model to the training data.

Prediction

To assess each model's ability to predict unseen data, we performed predictions 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 3 shows the fit of the test set data to the model predictions.

Discussion

From these results, we see 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

Table 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

Table 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		*	**	***

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.

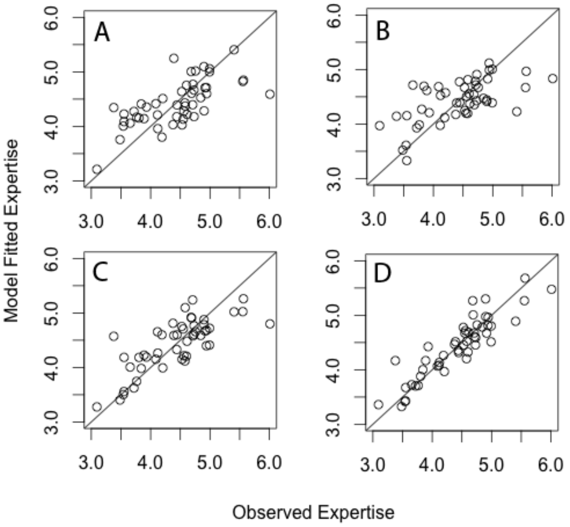


Figure 2. Fit of multiple regression model to training data. Different plots for models sampling from A) 2 episodes, B) 10 episodes, C) 100 episodes, and D) all observed episodes per game.

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, we see that 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 our pursuit of using small proportions of data to predict long-term performance.

It is tempting to draw conclusions from the lists of significant predictors presented in Table 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). We cannot yet account for these covert effects and are not prepared to draw strong conclusions about the individual predictors' viability in predicting long-term Tetris performance. We can, 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 categories (global, local, and immediate) tend to emerge as significant across samplings, indicating that there

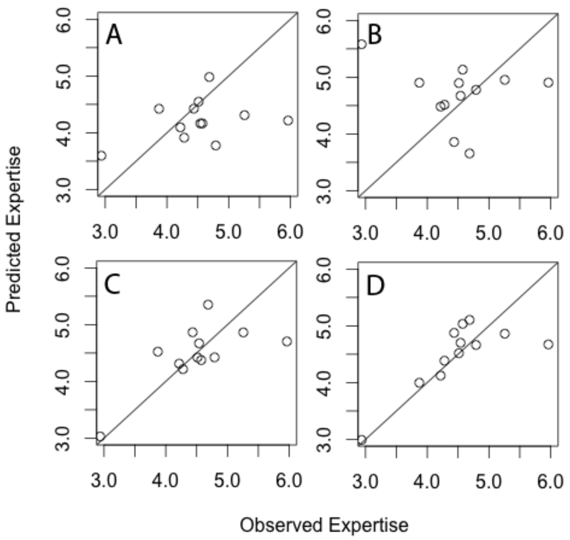


Figure 3. Fit of predictions from models to the test data set. Different plots for models sampling from (A) 2 episodes, (B) 10 episodes, (C) 100 episodes, and (D) all observed episodes per game.

may exist latent factors within each of these categories which contribute independently to skilled performance.

Conclusions

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

Although our results are tentative, we are pleased with our initial success in applying a general cognitive task approach to extreme expertise. Our 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 our initial items for each scale shows some success as a predictor of expertise. Thin-slicing seems to produce valid predictions as, to our surprise, even the regression model based on the first two episodes of each game had some predictive validity. We are embolden by these initial successes and have made plans to collect an order of magnitude more data from an order of magnitude more players at all levels of expertise.

Our predictive modeling has thus far been limited to the statistical technique of multiple regression. Other techniques have been suggested and we are openly soliciting suggestions from the cognitive community. 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.

Acknowledgements

We thank 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.

References

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*, 1–7. 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.

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*, 97–115.