Title
The Role of Variety in the Transfer of Cognitive Skills

Author
Bainbridge, Katie

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Santa Barbara

The Role of Variety in the Transfer of Cognitive Skills

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychological and Brain Sciences

by

Katie M. Bainbridge

Committee in charge:

Professor Richard E. Mayer, Chair

Professor Mary Hegarty

Professor David Sherman

Professor Tommy Sprague

September 2019

The dissertation of Katie M. Bainbridge is approved.

Mary Hegarty

David Sherman

Tommy Sprague

Richard E. Mayer, Committee Chair

August, 2019

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Richard E. Mayer, for giving me a strong foundation upon which to build my future academic career. I'd also like to thank my husband, Dr. Kyle Ploense, for his support, both practical and emotional, during these final stages of manuscript writing. A thank you is also owed to friends, such as Jonathan Ying and Chelsea Brown, for helping to feed me and my family while time and energy was short. Additionally I'd like to thank my parents, Kelli and Dwight Bainbridge, as well as my grandmother, Libby Perrin, for being marvelous caregivers to our daughter, Rosemary, periodically over the course of this final summer. Finally, I'd like to thank the undergraduate research assistants who made the data collection process feasible; in particular Jake Bassin, who was my faithful assistant throughout all four years of his undergraduate career.

VITA OF KATIE M. BAINBRIDGE
August 2019

EDUCATION

University of California, Santa Barbara | PhD | *2017*

- Department of Psychological and Brain Sciences
- Dissertation Title: “The Role of Variety in the Transfer of Cognitive Skills”
- Advisor: Dr. Richard E. Mayer

University of California, Santa Barbara | Master of Arts | *2017*

- Department of Psychological and Brain Sciences
- Thesis Title: “Cognitive Consequences of Short and Long Term Training with Lumosity Games”
- Advisor: Dr. Richard E. Mayer

Sarah Lawrence College, Yonkers, NY | Bachelor of Arts | *2011*

- Degree in Liberal Arts with focus on Neuroscience and Cognitive Development
- Advisor: Dr. Adam Brown

PUBLICATIONS

Mayer, R.E., Parong, J., **Bainbridge, K.** (2019) Young Adults Learning Executive Function Skills by Playing Focused Video Games. *Cognitive Development, 48,* 43-50

Bainbridge, K. & Mayer, R. E. (2018). Shining the light of research on Lumosity. *Journal of Cognitive Enhancement*, *2*(1), 43-62.

Bainbridge, K. & Brown, A. (2014). rTMS as a treatment for anorexia nervosa. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation, 7*(1), 149-150.

RESEARCH EXPERIENCE

UCSB | Santa Barbara, CA | *2014 - Present*

- PhD Candidate
- Learning and Instruction Lab
- PI: Dr. Richard E. Mayer
- Primary duties: Read background literature and propose new experiments regarding educational games, instructional technology, and learning sciences; design studies and create materials and measures using Direct RT, e*prime, and HTML; analyze data using SPSS and R; interpret and communicate findings during area meetings; write papers for publication; and present at conferences

Rutgers University | Newark, NJ | *2013-2014*

- Research Assistant at the Center for Molecular and Behavioral Neuroscience
- PI: Dr. Mark Gluck
- Primary duties: analyzed data for studies on implicit learning, compiled and synthesized literature review manuscript, coordinated with surgeons and Parkinson's clinics for upcoming deep brain stimulation project, and administered cognitive interviews for Parkinson's patients

John Wiley & Sons | Hoboken, NJ| *2011-2013*

- Bibliometric Data Analyst
- Primary duties: analyzed citation data for academic publishing, tracked trends across disciplines and geographical locations, cleaned data sets, and visualized data for management

Columbia University | New York, NY| *2011*

- Research Assistant at the Division of Brain Stimulation and Therapeutic Modulation
- PI: Dr. Joshua Berman
- Primary duties: Assisted with study on the efficacy of Electro Convulsive Therapy (ECT) for treatment-resistant depression, conducted clinical intake (blood samples, urine samples, EKG) for inpatient stay, learned about Transcranial Magnetic Stimulation (TMS), and proposed new directions for TMS research, culminating in a paper on its potential use in anorexia nervosa

The New School for Social Research | New York, NY| *2011*

- Research Assistant
- PI: Dr. Adam Brown
- Primary duties: Assisted with studies 1.) how group recall affects flashbulb memory accuracy and 2.) the relationship between memory and imagination

New York University| New York, NY| *2011*

- Research Assistant at the Department of Cognition and Perception
- PI: Dr. Marjorie Rhodes
- Primary duties: Recruited and scheduled participants for a study on the development of social stigma in preschool children and attended and contributed to lab meetings

TEACHING EXPERIENCE

California State University Channel Islands, Department of Psychology

Lecturer *2019 - Present*

- Course title: Research Methods
- Designed course materials, gave lectures, guided students through the research process as they developed and executed their own studies during lab, and assessed student work

University of California Santa Barbara, Department of Psychological and Brain Sciences

Instructor of Record *2017, 2018*

- Course Title: Educational Psychology
- Designed course, wrote and delivered lectures, tests, and quizzes, and facilitated discussions

Class Coordinator *2017 - 2018*

- Course Titles: Advanced Research Methods
- Updated and disseminated lab materials, led weekly meetings with all TAs, trained and supervised new TAs, attended all sections during the study design and data analysis phases to facilitate and advise groups, and managed grade book

Section Teaching Assistant *2014 - Present*

- Course titles: Advanced Research Methods and Statistics for Psychology
- Led my own sub-section of the lecture for classrooms of ~20 students
- Gave lectures, led discussions and lab activities, and graded written work

Administrative Teaching Assistant *2014 - Present*

- Course titles: Cognitive Neuroscience, Cognitive Psychology, Educational Psychology, Psychopharmacology of Drug Abuse, Brain States, and Introduction to Biopsychology
- Attended lectures, held office hours to assist with student understanding, and led review sessions prior to testing
- Assisted with administrative tasks such as managing the course website, keeping the grade book, and grading tests and quizzes

SERVICE

Pedagogy Consultant

Department of Instructional Development, UCSB | *2018 - Present*

- Consulted Teaching Assistants in all departments at UCSB on how to improve their teaching, conducted teaching workshops

TAAP: Teaching Assistant Advisory Panel

UCSB | *2015 - 2018*

- Mentored incoming first year graduate students in how to be teachers as part of a year-long Teaching Psychology course

Mentorship of Undergraduate Research Assistants

UCSB| *2014 - Present*

- Mentored undergraduate students who volunteered in my lab in experimental methods

- Taught students how to program tasks in Direct RT and e-prime, design experiments, recruit and schedule participants, follow an experimental protocol, enter data, and interpret statistical tests in the context of our research
- Total of seventeen students to date, including:
  - Students recruited from my classes: Nicole Bouette, Evan Bradshaw, Corey Brown, Rosaura Contreras, Kaylee Cortopassi, Jenalyssa Cruces, Andres Diaz, and Luis Garcia
  - Students recruited from wider PBS Department: Jake Bassin, Bianca Dalangin, Jana Miclat, Matthew Mihelic, Ixchel Morfin, Melissa Powell, and Pia Von Strasser
  - Students recruited through the Society of Undergraduate Psychologists (SUP) mentorship program: Jiajing Yan and Vanessa Veloz

INVITED TALKS AND PRESENTATIONS

Talks

Lightning Talk - Graduate Student Teaching Symposium | UCSB | *April 26, 2019*
"Why that funny cartoon in your slides isn't helping anyone: the coherence principle in multimedia learning"

Instructional Development Workshop | UCSB | *April 22, 2019 & March 15, 2019*
"Evidence-based PowerPoint design"

Lightning Talk Competition - Grad Slam| UCSB | *April 10, 2019*
"Can video games make you smarter?"

Instructional Development Workshop | UCSB | *April 1, 2019*
"Setting the tone on the first day of class"

Instructional Development Workshop | UCSB | *February 5, 2019*
"Handouts: 101"

Posters

Cognitive Science Society Meeting | Poster Presentation | *Montreal, 2019*
"Implicit and Explicit Strategies in Video Game Training"

APS Conference | Poster Presentation | *Washington DC, 2019*
"Role of Variety in Cognitive Improvement after Video Game Training"

Psychonomic Society Meeting | Poster Presentation | *New Orleans, 2018*
"Adding Variety to Mental Rotation Skill Training"

APS Conference | Poster Presentation | *San Francisco, 2018*
"Learning from a PowerPoint vs. a Google Search"

ABSTRACT

The Role of Variety in the Transfer of Cognitive Skills

by

Katie M. Bainbridge

Brain-training games are largely ineffective at improving player performance on cognitive measures, whereas action video games are largely effective at doing so, despite not being designed for that purpose. This dissertation draws upon the schema induction literature to propose that greater variety during the training phase is responsible for this discrepancy. Three studies training three different cognitive domains (mental rotation, working memory, and visual attention) were conducted comparing a condition that trained with relatively low variety of stimuli or environments to a condition that trained with relatively high variety. In Experiment I a group that trained mental rotation using three categories of shapes was compared to three groups that trained mental rotation with just one of the shape types on near, medium, and far transfer measures of mental rotation and spatial visualization. No difference was found between the low variety groups and the high variety groups on any of the cognitive posttest measures. In Experiment II a group that trained with an n-back using just letter stimuli was compared to a group that trained with an n-back using 6 types of stimuli and an inactive control group on near, medium and far transfer measures of working memory and fluid intelligence. The high variety group did not improve relative to the low variety group, and in some cases the low variety group improved more than the high variety group. In Experiment III participants with minimal prior experience with video games were assigned to play *Call of Duty* in a high variety condition, a low variety condition, or an inactive control condition. These groups were compared on their improvement on three measures of visual attention skill. In no case did the high variety group improve more than the low variety group. When the active groups were combined they did not improve more than the inactive control group. Across all three studies no benefit was found for the high variety condition relative to the low variety condition on related transfer measures of the respective cognitive skills. It is concluded that variety is not a useful addition to training interventions if the goal is to increase low-road transfer of cognitive skills.

TABLE OF CONTENTS

Chapter I

General Introduction

Research Question

It is a hope of many educators and companies alike that videogames may be able to improve cognitive performance (Gee, 2007; Hardy, Drescher, Sarkar, Kellett, & Scanlon, 2011; McMonigal, 2011; Prensky, 2006; Shaffer, 2006; Squire, 2011). While there is some promising evidence that playing action video games can improve perceptual attention skills related to gameplay (Bediou, Adams, Mayer, Tipton, Green, & Bavelier, 2018; Green & Bavelier, 2003; Green & Bavelier, 2015, Shute, Ventura, Ke, 2015; Mayer, 2014), endeavors to design a commercial brain-training game have largely proven ineffective at transferring improvements to tasks outside of the game (Bainbridge & Mayer, 2017; Simons, Boot, Charbness, Gathercole, Chabris, et. al., 2016; Hulme & Melby-Lervåg, 2013; Lorant-Royer, Spiess, Goncalves, Lieury, 2008; Mayer, 2014). Why would the action video game *Medal of Honor* be effective at improving cognition (Green & Bavelier, 2003) when the brain-training game *Lumosity* is not (Bainbridge & Mayer, 2017)? What are recreational action video games, which are not intended to improve cognition, doing on accident that commercial brain-training games, which intend to improve cognition, are failing to do?

The studies described below are designed to investigate *the variety hypothesis* as an explanation for this disparity. This hypothesis posits that greater variety in surface features during implicit practice leads to more flexible schemas that are more easily transferred to new environments. In this framework, action video game players must apply their skills more variably than brain-training game players which leads to an improved ability to abstract the skill to new contexts and better transfer. The greater the variety experienced during learning, the greater the likelihood the acquired skill will transfer.

Literature Review

The problem with transfer. Transfer is the phenomenon by which learning acquired in one context is successfully applied in an unrelated context. A person might learn to calculate percentages in school, and then that knowledge transfers to calculating the tip on a bill at a restaurant. We know that transfer must occur; if it did not humans would be unable to function in unfamiliar situations. However, transfer is notoriously hard to observe in a laboratory setting (Barnett & Ceci, 2002; Bransford & Schwartz, 1999; Cooper & Sweller, 1987; Day & Goldstone, 2012; Singley & Anderson, 1989). A potential explanation for this difficulty is that there are multiple kinds of transfer. One framework, as described by Saloman and Perkins (1989) breaks transfer into two main paths: high-road transfer (which occurs when you consciously recognize that learning from a previous situation applies to a new one) and low-road transfer (which occurs unconsciously when learning from a previous situation is spontaneously used in a novel situation without intent). It is more likely that cognitive skills, such as visual attention, transfer via the low-road route, as the process is not deliberate.

Evidence that action video games work. A comprehensive review of studies investigating the effect of action video games found converging evidence both for the notion that habitual action video game players demonstrate improved performance on cognitive measures compared to non-gamers, and that long-term interventions that require non-gamers to play action video games improve performance on cognitive measures related to skills used in gameplay (Bediou, et. al., 2018). The effects are mostly limited to studies that use first person and third person shooter games, and effect sizes are somewhat larger for comparisons of habitual gamers to non-gamers (g = 0.55) than for intervention studies that attempt to induce changes with training (g = 0.34). Likewise, a review by Mayer (2014) of action video game literature revealed consistent improvements on visual attention measures in studies that had participants train with first person shooter games with effect sizes ranging from d = 0.26 to 2.61.

One of the most famous examples of the effects of action video game training comes from Green and Bavelier (2003). They demonstrated that habitual action video game players far exceeded non-gamers on measures of visual attention and spatial skills. They acknowledged that this result could reflect a selection bias, wherein people naturally better at these skills are potentially more drawn to a hobby that utilizes them. To account for this, they had non-video-game-players play *Medal of Honor,* a first-person shooter action video game, for 1 hour a day for 10 consecutive days and compared them to non-gamers who played *Tetris* for the same training period. They found that the group that played *Medal of Honor* outperformed the *Tetris* group on all three measures of relevant cognitive skill transfer that they used.

One study compared an action video game to a brain training game directly (Shute, Ventura, Ke, 2014). One group played *Portal* for 8 hours and another played *Lumosity* brain-training games for 8 hours. Before and after training, participants took measures of problem solving, spatial ability, and persistence. On all three measures, participants who played *Portal* improved more than participants who played *Lumosity,* despite the fact that *Lumosity* claims to improve cognition and *Portal* makes no such claims and is intended purely as a recreational game.

Evidence that brain-training games do not work. A collection of 75 cognitive scientists signed a statement in 2015 asserting that the evidence supporting brain-training games was thin ("A Consensus on the Brain Training Industry from the Scientific Community", 2015). Indeed, meta-analyses and well-controlled experimental studies have failed to produce evidence supporting the notion that playing brain-training games can improve cognitive skills. A meta-analysis of all experiments using CogMed training, a brand of brain-training games, conducted by Hulme and Melby-Lervåg (2013) concluded that training with CogMed doesn't improve any of the cognitive measures it claims to improve, including attention, vigilance, or working memory. Likewise, a more general meta-analysis of brain training through games and cognitive exercises by Simons et. al. (2016) concluded that while brain-training interventions can improve the trained task, there is almost no evidence that this improvement transfers to untrained tasks. A review by Mayer (2014) concluded that the evidence for brain training games was mixed, and what evidence did support their efficacy had a small median effect size.

An experiment by Lorant-Royer, Spiess, and Goncalves (2008) compared children that played brain-training games, children that did paper-and-pencil training exercises, and an inactive control group. They found that children who played the games improved on two of six cognitive measures, but the paper-and-pencil group and inactive control group also improved on one of those measures. In addition to cognitive measures they also analyzed improvement on three academic measures and found no evidence of transfer to these measures after training. They conclude that brain-training games are not effective.

One study attempted to evaluate the efficacy of *Lumosity* brain-training games in a way that directly tested their claims (Bainbridge & Mayer, 2017). The training schedule selected was identical to the one recommended by *Lumosity*, i.e.15 minutes a day for 5 days a week for 16 weeks, totaling in 20 hours of training. The cognitive measures selected were specific to the domains the games selected claimed to improve. One group played 5 games designed to improve visual attention, one group played 5 games designed to improve flexibility and inhibition, and one group served as an inactive control. All participants took 2 measures of visual attention and 2 measures of flexibility/inhibition both before and after the training period. If *Lumosity's* claims were correct, the participants who trained with attention games should have improved on the attention measures relative to the flexibility group and the control group, and the group that trained with flexibility games should have improved on flexibility measures relative to the attention group and the control group. Instead, the researchers found improvement for only one active group on only one task, and the task they improved on was almost identical to a game included in the training. They conclude there is no evidence that training with *Lumosity* transfers meaningfully to cognitive improvement on the targeted skill outside of the games (Bainbridge & Mayer, 2017).

Evidence that variety is important for transfer. Outside of the realm of game research there are numerous lines of study that suggest variety may be an important aspect of whether learning will transfer to a novel situation. For example, Gick and Holyoak (1983) had participants read a story about a general who wanted to attack a fortress. The bridges leading to the fortress had mines that would go off if a significant ground force crossed them, but would not go off if just a few people at a time crossed. The general's force was large enough to take the fortress, but was too large to cross any one bridge without setting off the mines. Instead he divided up his troops between the bridges to converge on the fortress from different directions. The participants were then presented with Dunker's radiation problem: A patient has an inoperable tumor. A doctor knows he can destroy the tumor, but with the amount of radiation needed to destroy the tumor he would also destroy any healthy tissue along the way. A lower level of radiation would not kill the healthy tissue, but it would also fail to kill the tumor. The participants must come up with a solution that kills the tumor without killing the healthy tissue. Both the story and the problem require a convergence solution: The doctor should use a lower level of radiation and attack the tumor from multiple angles, just like the general dividing his force between the bridges. The surface details are different, but the underlying structure of the problem and solution are the same.

When presented with just the radiation problem, 10% of participants come up with the convergence solution spontaneously. If participants read the general story immediately before the radiation problem, the solution rate only goes up to 30%, suggesting only 20% of participants use the general problem to solve the radiation problem. When participants are explicitly told that the general story can help them solve the radiation problem, the solve rate jumps to 75%, but that still leaves 25% of the participants who fail to transfer the solution from one story to the other. Gick and Holyoak tried multiple methods for increasing the solution rate: They had participants summarize the general story before reading the radiation problem, they provided participants with the stripped-down principle of the general story before they read the radiation problem, and they provided a visual diagram illustrating the convergence solution with the general story. None of these interventions improved the solution rate. The only technique that did increase transfer from the general story to the radiation problem was adding a second dissimilar story with the same structural features before presenting participants with Dunker's radiation problem. They conclude that seeing the convergence solution presented in multiple dissimilar situations helped them notice its applicability when presented with a third dissimilar situation (Gick & Holyoak, 1983).

Posner and Keele (1968) were investigating schema acquisition when they also stumbled upon the principle that variety in learning improves transfer. They showed participants dot patterns in a matrix generated from a prototypical pattern. One group saw patterns with minor distortions from the prototype, the other group saw patterns with high distortions from the prototype. They hypothesized that if acquiring an accurate schema was more important for learning transfer, the group shown minor distortions would be better able to categorize novel dot patterns, whereas if variety was more important for learning transfer, the high distortion group would perform better. After their training, both groups were shown new patterns and were asked if they belonged to the same category they had learned. The high distortion group was better at categorizing these patterns, suggesting that variety during learning was more important than an accurate schema when transferring learning to novel stimuli.

Another series of experiments that corroborates the importance of variety comes from Chen and Mo's (2004) revisit of Luchins' water jar problem experiment. Luchins famously illustrated that learning one formula for getting the desired volume of water in one jar led to negative transfer; participants were more likely to use that formula, even when an easier solution was available (Luchins, 1942). Chen and Mo similarly asked participants to get from one starting quantity to another using an algebra formula, but for one of the groups they varied the practice problems across multiple dimensions. They found that varying the surface features of the practice problems led to slower learning of the procedures, but more flexible schemas that more easily transferred to new problems.

Theory

At this point it may be useful to revisit the concepts of low-road transfer and high-road transfer (Salomon & Perkins, 1989). Low-road transfer is implicit, automatic, and non-conscious. It reflects extended practice, and is acquired by using target skills in variable contexts. The distance the learning will transfer depends on the amount and the variability of the practice. Salomon and Perkins liken low-road transfer to learning to drive lots of different cars, each with different gear shifts and steering wheels and clutches, and then attempting to drive a tractor trailer. If you'd only ever driven one car, transitioning to the tractor trailer would likely be more difficult than if you were accustomed to switching your driving knowledge between multiple cars. High-road transfer, on the other hand, is conscious and explicit. One only needs a single exposure to a base problem that is applicable to a target problem in order for it to occur, but recognizing the connection between the problems consciously can be difficult. The distance of the transfer depends on one's ability to recognize the applicability of a solution or procedure one has already learned to a novel context.

Using the general story to solve Dunker's radiation problem as described in the Gick and Holyoak (1983) study above is an example of high-road transfer. The pattern recognition described in the Posner and Keele (1968) above is an example of low-road transfer. The Chen and Mo (2004) experiments likely draw on both low-road and high-road transfer. In both types of transfer, learning the target skill with variable surface features will increase the likelihood of it being activated in a novel context, but this is particularly important in low-road transfer, as the recognition is automatic and non-conscious. In high-road transfer it is possible to recognize a base problem after just one exposure, but it is still easier with more variety in learning. Low-road transfer is probably the more relevant mechanism to improving cognitive performance with both action video games and brain-training games. It is unlikely that habitual action video game players like the ones in the Green and Bavelier (2003) study are consciously and deliberately applying the visual search skills they practiced while playing first-person shooter games to new contexts. The improvement more likely represents an automatic activation of skills that have been developed with implicit practice in highly variable contexts.

Related to low-road transfer is the concept of implicit practice, as described by Uttal and Meadow (2013). When learning occurs through implicit practice, one is not consciously aware that the learning has taken place. The learner is not directing their learning towards some goal. As an example, Uttal and Meadow describe the crowding phenomenon. When visual stimuli are presented very close together they become harder to discern than when they are presented further apart. A habitual video game player demonstrates a smaller crowding region; they don't suffer as much from crowding effects and are able to discern objects that are presented close together with greater speed and accuracy than non-video game players. The video game players did not seek out this skill, and they are often not consciously aware of how it is benefitting them. It arose naturally as a result of implicit practice (Uttal & Meadow, 2013). As in the low-road transfer example of driving many different cars, the learning acquired by switching vehicles is a form of implicit practice. The varying nature of the surface differences between vehicles will give the learner an unconscious, automatic *feel* for the structural similarities, and they will more easily apply that learning when asked to drive a new, dissimilar vehicle.

Both low-road transfer and implicit practice are related to schema acquisition as described in the Posner and Keele study (1968) described above. While schemas can be conscious and deliberate, the schemas they induced arose naturally and automatically through implicit practice. If the experiences that create the schema are very tightly related, the schema will be well-defined but narrow. It will apply very well to similar cases but will not be easily applied to dissimilar cases. If the implicit practice is highly varied and includes dissimilar surface features, the schema will be more flexible. It may not be as easily or quickly applied to archetypical cases as the narrow schema, but it will more easily be applied to novel cases. If the schema was developed under conditions where surface details were consistent and reliable, a related situation where the surface details are unfamiliar will likely result in a failure to transfer. If the schema developed under conditions where the surface details were varied and could not be relied upon, a transfer case with unfamiliar surface features but similar structural features would not hinder the application of the schema. It was developed in a flexible environment and can be applied flexibly.

In the low-road transfer car analogy, a person who has only ever driven the same car who is suddenly asked to drive a tractor trailer would likely be stumped by the differences. They have only ever had to reach for the same location when turning on the windshield wipers, and now both the location and mechanism for doing so are completely different! In contrast, a person who has driven a variety of cars would have faced that problem repeatedly. They would not be caught-up in the surface differences, and would instead notice the structural similarities. The windshield wipers in some cars are on the right side and sometimes on the left side of the steering wheel. Sometimes they are operated with a twist and other times with a button. When driving the unfamiliar tractor trailer as it starts to rain their implicit learning from the various other vehicles will make identifying and activating the windshield wipers a more automatic, less effortful process than the one-car-driver experienced. Both could have the same number hours of driving experience, but the flexibility of their schemas would be more or less easily transferred as a function of the variability of that practice.

From the studies discussed in the literature review above we can conclude that improving cognitive performance through a game is possible, but that brain-training games are failing to do so. This dissertation proposes the variety hypothesis, which suggests that the amount of variety experienced during training can explain this discrepancy. Greater variety in the learning environment leads to more flexible schemas that can more easily be applied to novel situations via low-road transfer. In an action video game a player would be asked to apply the same skill repeatedly, but almost never in the same context twice. For example, in a first-person shooter game like *Medal of Honor* or *Call of Duty,* a player would need to visually scan the environment constantly for targets and respond quickly and appropriately to those targets, but the environment itself and the appropriate response would be consistently different. In contrast, a brain-training game like *Lumosity* would also ask a player to apply the same skill repeatedly, but the context would be remarkably consistent, with only a handful of different environments. A certain amount of player improvement could be attributed to environment-specific, explicit strategies that would be of no use outside of the game context. In an action video game, there is no point in investing in these environment-specific strategies, as the environment changes too rapidly for them to pay off.

If this proposed role of variety in transfer is correct, training tasks that demand subjects apply the target skill in more varied ways will lead to better transfer performance on related cognitive skills than training tasks that have subjects apply the target skill in the same environment with the same surface features repeatedly.

Operationalization

The following studies seek to investigate this claim that greater variety during a learning phase leads to a more flexible implicit schema that is more easily transferred to a novel situation.

Each subsequent experiment increases the amount of variety and amount of training. The first experiment trains mental rotation skill for one hour. Variety is operationalized by increasing the types of shapes with which the participants practice. Three of the groups practice mentally rotating just one type of shape. They are compared to a fourth group who practices mentally rotating all three of the shape types. They are assessed on near transfer (using a nearly identical task with novel shape types), medium transfer (using more dissimilar mental rotation tasks), and far transfer (using spatial visualization tasks).

The second study switches domains to train working memory ability for 5 hours. Variety is operationalized by comparing working memory training with just letter stimuli to working memory training with 6 dissimilar types of stimuli. The participants are assessed on near transfer (using a nearly identical working memory task with novel stimuli), medium transfer (using other, unpracticed working memory tasks), and far transfer (using a fluid intelligence task). An inactive control group is added to account for practice effects.

The final study applies the principle of variety directly to action video game training. Variety is operationalized by comparing a group that plays an action video game in the same environment for 9.5 hours to a group that plays the same action video game for the same amount of time in a different environment every day. Both groups are compared to an inactive control group to account for practice effects. Transfer is assessed with improvement from pretest to posttest on three different measures of visual attention skill.

Variety Hypothesis

Two main predictions, which I will call the variety hypothesis, remain constant throughout all three studies, regardless of whether the participants are training mental rotation, working memory, or visual attention:

1) The group that trains with a higher amount of variety (referred to henceforth as the High Variety group) will not perform as well during the learning phase as the group that trains with a limited variety (referred to henceforth as the Low Variety group). This relates to the schema development mechanism described above. The Low Variety group will develop a more specific, well-defined schema that is more useful during training. The High Variety group's schema will be less useful during the learning phase because it must account for the inconsistency of the surface details, but for this same reason will be more flexible and more easily applied in new situations. This hypothesis is bolstered by the findings from

Chen and Mo (2004), who found that more consistent surface features during learning benefitted performance during the learning phase.

2) The High Variety group will more easily transfer their learning to novel situations relative to the Low Variety group, and their transfer will extend further than the Low Variety group. The schemas they develop will be less affected by changes in surface features, and will therefore be more easily applied in un-practiced contexts. This hypothesis is also consistent with findings from Chen and Mo (2004) who found that the group that practiced with more varied surface features performed better on transfer problems, despite not doing as well as the group that practiced with consistent surface features during the learning phase.

If the variety hypothesis is supported across domains it would indicate that greater variety during learning that involves implicit practice is an important factor in whether that learning will transfer. This finding would tie together the training, learning transfer, and schema induction fields. It would also help to explain why the results of cognitive training interventions are so inconsistent, and would provide a lens through which researchers could judge why some training interventions are effective and others are not. The variety hypothesis could also be used to boost the efficacy of future game-training interventions, which could help bring the utopian vision of improving cognition by playing a game come to fruition.

Chapter II

Experiment I: The Variety Hypothesis in Mental Rotation Training

Objective

Spatial ability has been linked with increased interest and aptitude in Science, Technology, Engineering, and Math (STEM) fields (Newcombe, 2010; Stieff, 2011; Hegarty & Waller, 2005; Wai, Lubinski, & Benbow, 2009). Due to this link, it has been suggested by some that improving visuospatial reasoning could potentially improve STEM outcomes in students (Stieff & Uttal, 2015; Uttal & Cohen, 2012). Improvements in spatial ability would ideally be trained through low-road transfer techniques such as implicit practice as the ability feels automatic and non-conscious to those who employ it (Salomon & Perkins, 1989); however, intentional game-training interventions have been notoriously ineffective at conveying improvements to the players (Bainbridge & Mayer, 2017; Kable, et al, 2017; Simons, Boot, Charbness, Gathercole, Chabris, et. al., 2016; Hulme & Melby-Lervåg, 2013; Lorant-Royer, Spiess, Goncalves, Lieury, 2008; Mayer, 2014). This study seeks to test whether greater variety during mental rotation training improves transfer. We define *variety* as a greater amount of dissimilar surface features. In this case, this refers to differing shape types. If effective, this element could be added to future game training interventions to enhance outcomes.

Literature Review

In a meta-analysis of spatial training studies, Uttal and Meadow (2013) found that spatial skills respond strongly to training (d = 0.47). A second meta-analysis by Uttal and colleagues (2015) found that these improvements transfer to novel tasks with an effect size of d = 0.48. One of the most common methods of training spatial skills is through mental rotation practice (Uttal & Meadow, 2013; Uttal, et al, 2015). The classic paradigm for training mental rotation involves presenting participants with a stimulus, then asking them to judge whether that stimulus is a rotated or mirrored version of a target stimulus (Cooper, 1975; Shepard & Metzler, 1971). This paradigm is still used to this day, and can show improvements in as little as one hour of training (Adams, Stull, & Hegarty, 2014).

Two mental rotation training studies in particular highlight the importance of considering how much variety participants were exposed to when evaluating the success or failure of a training intervention. In one, Terlecki, Newcombe and Little (2008) had participants play either Tetris or Solitaire for one hour a week for 12 weeks. Both groups took the Shepard and Metzler Mental Rotation Task (MRT) once a week for each of these 12 weeks. Transfer tests of spatial ability were administered before the training, immediately after the training, and 2-4 months after the training. Both groups improved on the Shepard and Metzler MRT, but the Tetris group improved on the transfer tests compared to the Solitaire group both immediately after the training and at the 2-4 month follow-up. This finding stands in direct contrast to a study by Sims and Mayer (2001), who likewise had participants play Tetris for 12 hours; however, these participants did not improve on transfer tests relative to control participants. They only improved their ability to rotate Tetris shapes. One possible explanation for the difference between these results can be attributed to the repeated testing paradigm administered by Terlecki and colleagues (2008). By having their participants repeatedly take the Shepard and Metzler MRT, they introduced an element of variety to the spatial training. The Tetris participants in the Sims and Mayer (2001) study had no such variety in their training, hence the specificity of their improvement.

Theory and Predictions

As described in the general introduction, research from the domain of schema induction (Chen & Mo, 2004; Posner & Keele, 1968) indicates that learning environments with increased variety lead to poorer performance during the learning phase but better performance when applying the learning to new situations.

This dissertation proposes the variety hypothesis: that greater variety during the learning phase hurts performance during the learning task, but increases the likelihood that learning will transfer to novel situations. The proposed mechanism is tied to the schema induction literature: if the prototype learned is too rigid and developed from cases that are too closely related, the learner is less likely to recognize when a dissimilar case arises where the learning would still be applicable. If the learning phase encompasses a wider range of examples, the prototype developed is more flexible and more likely to be applied in situations that were not explicitly practiced during the learning phase, increasing the probability and distance of transfer.

As it applies to mental rotation training, this proposed mechanism is supported by findings from Tarr and Pinker (1989), who found that trained effects for mental rotation were limited to the figures and orientations specifically practiced. This contradicts the strong effect size for transfer found by the Uttal and colleagues meta analysis (2013). Perhaps the discrepancy can be explained by the variety hypothesis. If the shape type used for training is too limited, the strategy developed is too specific. If the training involves a high variety of shapes, the strategy of encoding specific shapes and orientations does not pay off for the learner, so they must invest in other strategies such as spatial visualization. Investing in this strategy would be a more challenging and effortful process, but would help them more in transfer tasks than the specific shape encoding strategy.

The proposed mechanism for why higher variety would result in greater transfer is that participants will not invest in an explicit, specific shape encoding strategy, as seen in Tarr and Pinker (1989) and will instead invest in an implicit spatial visualization strategy.

To test whether increased variety improves transfer after a mental rotation training intervention, we compared participants who train with just one shape type (Low Variety group) to participants who train with 4 shape types (High Variety group). It was hypothesized that:

1. The Low Variety group would perform better during the training task than the High Variety group, responding to trials with greater accuracy and faster reaction times.
2. The High Variety group would report more challenge, more effort, and more enjoyment on the training task than the Low Variety group.
3. The High Variety group would perform better on transfer tests of mental rotation (Card Rotation, Cube Comparisons, Shepard and Metzler Mental Rotation) than the Low Variety group, and will also perform better on far-transfer tasks of spatial visualization (Surface Development, Geologic Blocks) than the Low Variety group.

Method

Participants

The participants were 187 undergraduate psychology students at the University of California, Santa Barbara who participated in the study in return for class credit. In a between subjects design, 50 students were randomly assigned to the Tetris shape condition, 47 were randomly assigned to the Polygons condition, 41 were randomly assigned to the Chemistry shape condition, and 49 were randomly assigned to the High Variety condition which saw all shape types.

Materials

There were four training tasks, all of which were made using Direct RT: Tetris shape training task, Polygon training task, Chemistry shape training task (which make up the single shape training tasks) and a multiple shape training task. The pre and post tests were administered on paper or in the case of the Shepard and Metzler Mental Rotation Task with e-prime.

Training Tasks. Training consisted of 400 trials divided into four progressively more difficult blocks. There were four training groups: Tetris shapes, Polygon shapes, Chemistry shapes, and a group that trained with all shapes used in the previous three groups. When referenced collectively, the single-shape training groups will be called the Low Variety group, and the group that trained with all shape types will be called the High Variety group. During the practice phase all groups were presented with two shapes. They were asked to indicate if the shape on the right was a rotated or a mirrored version of the shape on the left. Each category (Tetris shapes, Polygons, Chemistry shapes) had 26 different shapes, but the degree of rotation for those shapes became progressively greater as the task went on. For the first 100 trials the shapes were rotated between 20 and 60 degrees. For the second 100 trials the shapes were rotated between 60 and 100 degrees. For the third 100 trials the shapes were rotated between 100 and 160 degrees. For the final 100 trials the shapes were rotated between 20 and 160 degrees. Between each 100 trial block participants were instructed to take a break and rest their eyes. In all conditions on all trials participants were given feedback. If they were correct, they were shown a screen that said “Correct!” with a smiling face. If they were incorrect, they were shown a screen that said “Incorrect” with a frowning face.

Tetris shape training task. Tetris shapes were two dimensional and solid black in color. Twenty six different Tetris shapes were included in the training. An example of a training trial using a Tetris shape can be seen in Figure 1.

Polygon training task. Polygons were two dimensional and solid black in color. There were 26 different polygon shapes; these were all irregular and did not include any common shapes such as squares or triangles. An example of a training trial using Polygons can be seen in Figure 2.

Chemistry shape training task. Chemistry shapes were two dimensional representations of molecules, solid black in color. They did not include any chemical notation, as this would have made it easier for participants to recognize of the shape was rotated or mirrored without using spatial visualization. An example of a training trial using Chemistry shapes can be seen in Figure 3.

***High Variety Training Task*.** This condition practiced mental rotation using Tetris shapes, Polygons, *and* Chemistry shapes. Which shape type was presented was randomly interleaved. Everything else about the training task, including the instructions, number of trials, degree of rotation in each block, and trial feedback, was the same as the single shape training tasks.

Dependent Measures. The dependent measures involved a questionnaire and five spatial ability tests. The Card Rotation test, Cube Comparison test, Surface Development test, and Geologic Block Cross Sectioning test were administered on paper, and the Shepard and Metzler Mental Rotation test was administered using e-prime. Spatial ability was measured after training using 5 posttests ranging from near transfer (two-dimensional mental rotation, which is most closely related to the training task, and was captured by the Card Rotation test) to moderate transfer (three-dimensional mental rotation, which is somewhat related to the training task, and was captured by the Cube Comparison and Shepard and Metzler Mental Rotation tests), to far transfer (spatial visualization, which is least related to the training task but still related to the target skill, and was captured by the Surface Development and Geologic Block Cross Sectioning tests).

Questionnaires. All participants were given a questionnaire that assessed their enjoyment of the task and how challenged they felt by the task. All questionnaire items used a 1-5 Likert scale with 1 being "Strongly Disagree" and 5 being "Strongly Agree". An example of an enjoyment item is "I would like to train with more tasks like this". An example of a challenge item is "Please rate how much effort you exerted in this training task". The full questionnaire can be viewed in Appendix A.

Card Rotation Test. This test was taken from the Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman & Dermen, 1976) and was administered on paper. Participants are shown a target two-dimensional shape followed by a series of rotated versions of that shape. Participants have 3 minutes to indicate on as many trials as they can if the rotated shapes are the same or a mirrored version of the target shape. Half of the Card Rotation Task (10 items in 3 minutes) was given before the training begins to assess baseline 2D mental rotation ability. Half of the Card Rotation Task (10 items in 3 minutes) was given after the training to assess improvement. This task represents the nearest level of transfer between the training tasks and the assessment tasks. For sample items, see Figure 4.

Cube Comparison Test. This test was taken from the Kit of Factor-Referenced Cognitive Tests (Ekstrom, et.al., 1976) and was administered on paper. Participants are shown two cube shapes with letters drawn on the three visible sides. They have 6 minutes to judge whether the two cubes are the same or different based on the position of the letters. This task represents a moderate level of transfer between the training task and the assessment tasks. For sample items, see Figure 5.

Shepard and Metzler Metal Rotation Test (MRT). This is a computerized version of the classic Shepard and Metzler task and was administered via e-prime. It shows participants a target 3D shape followed by a series of rotated or rotated and mirrored versions of that shape. They are asked to judge whether the successive shapes are the same or mirrored from the target shape (Shepard & Metzler, 1988). This task represents a moderate level of transfer between the training task and the assessment tasks. For sample items, see Figure 6.

Surface Development Test. This test was taken from the Kit of Factor-Referenced Cognitive Tests (Ekstrom, et.al., 1976) and was administered on paper. It shows participants an unfolded version of a 3D object with each side marked with a number and are then shown the folded version of that object with each side marked with a letter. They are asked to pair as many of the numbered side with the correct lettered side as they can in 6 minutes. This task represents a far level of transfer between the training task and the assessment tasks. For a sample item, see Figure 7.

Geologic Block Cross Sectioning Test. This test was adapted from a task by Ormand and colleagues and was administered on paper (Ormand, et.al., 2014). It shows participants a 3D geologic block diagram along with a hypothetical sagittal slice that will be made in that block. They are then shown a series of potential 2D views if the interior of that geologic block. They must select which view they would see if that hypothetical sagittal slice were made in the 3D version of the block and complete as many of these trials as they can within 5 minutes. This task represents the farthest level of transfer between the training task and the assessment tasks. For a sample item, see Figure 8.

Procedure

Participants came to the lab and were given a consent form. In order to establish baseline mental rotation ability while still limiting the amount of variety they were exposed to before the training they were given the first 10 items of the Card Rotation Task, which took 3 minutes. They were then randomly assigned to one of 4 groups: the Tetris shapes group, the Polygons group, the Chemistry shapes group, or the High Variety group, which viewed all shape types. Tetris shapes, Polygons, and Chemistry shapes served collectively as the Low Variety group. The participants were then asked to complete 400 trials, which took approximately 1 hour, of the training task in their respective conditions. Each trial presented the participant with two versions of the same shape. The participant was asked to indicate whether the shape on the right was a mirrored or rotated version of the shape on the left. Every 100 trials the participant was given an opportunity to rest their eyes, and then the degree of rotation, and thus the difficulty of the task, would increase for the next block. After the 400th trial the participants were given a longer break. As part of this longer break the participants completed the Questionnaire and a distractor task. Collectively this long break took about 20 minutes. After the long break participants completed the second half of the Card Rotation Task, the Shepard and Metzler Task, the Surface Development Task, and the Geologic Block Cross Sectioning Task. Collectively these posttests took around 40 minutes. Upon completion, participants were thanked, debriefed, and released. The total time of the study was between an hour and a half and two hours. The study had IRB approval and was conducted in adherence with guidelines for treatment of human subjects.

Results

In order to compare performance across groups and measures, performance on pretest and posttest measures were standardized. Standardized posttest measures were averaged together across near, moderate, and far transfer categories. To make comparisons between high and low variety conditions clearer, all low variety conditions (Tetris shapes, Polygons, and Chemistry shapes) were combined into a single group (“Low Variety group”). Table 1 shows the collapsed means and standard deviations. Table 2 shows the unstandardized means and standard deviations on all measures for all groups separately.

Pretest Differences

An analysis of variance on card rotation pretest accuracy revealed no difference among the groups, $F(3, 183) = 1.17$, $p = 0.32$, indicating that they did not differ on mental rotation ability before the training. Means and standard deviations for pretest performance can be seen in Table 2.

Performance on Training Task

High Variety vs. Low Variety. Hypothesis 1 posited that the High Variety group would perform worse on the training task than the Low Variety group. Indeed, the Low Variety group was more accurate during training ($M = 92.81$, $SD = 5.99$) than the High Variety group ($M = 89.66$, $SD = 6.29$), $t(183) = 3.09$, $p < 0.01$. The Low Variety group ($M = 2042.94$ms, $SD = 926.71$) also had faster reaction times than the High Variety group ($M = 2471.60$ms, $SD = 767.86$) throughout the training, $t(185) = -2.90$, $p < 0.01$. These results support the hypothesis that the Low Variety group would do better during the training than the High Variety group.

All groups. When all four groups are analyzed separately, an ANOVA shows a significant difference between groups on training accuracy, $F(3,181) = 25.02$, $p < 0.00$, and training reaction time, $F(3, 183) = 22.56$, $p < 0.00$. Tukey post hoc tests reveal that on both training accuracy and reaction time, the Tetris shapes group performed significantly better than the Polygon group and the High Variety group, but not the Chemistry shapes group. The Polygon and High Variety groups performed significantly worse than the Tetris group and the Chemistry group but were not significantly different from each other. And the Chemistry group performed better than all other groups on both accuracy and reaction time during the training task. Table 2 shows the means and standard deviations for all training conditions.

Questionnaire Responses

The second hypothesis suggested that the poorer performance during training would be the result of the High Variety group being challenged more and needing to exert more effort during the training than the Low Variety group. As hypothesized, the Low Variety group reported significantly less challenge than the High Variety group, $t(185) = -2.09$, $p = 0.04$. Likewise the Low Variety group reported significantly lower effort than the High Variety group during the training, $t(185) = -2.26$, $p = 0.03$. It was also hypothesized that the High Variety group would enjoy the training more than the Low Variety group, but they did not report significantly different levels of enjoyment, $t(185) = -1.01$, $p = 0.30$. The hypothesis that the High Variety group would experience more challenge and effort than the Low Variety group was supported, but the hypothesis that the High Variety group would enjoy the training more than the Low Variety group was not supported.

Performance on Posttests: High Variety vs. Low Variety

Hypothesis 3 posited that the High Variety training would demonstrate better performance on transfer measures compared to participants who received the Low Variety training. The posttests were divided into three levels: near, moderate, and far transfer, with near transfer representing tasks most similar to the training and far transfer representing tasks most dissimilar from the training while still utilizing the target skill. Performance on the posttests was standardized and averaged across the transfer category. Table 1 shows the means and standard deviations for these analyses.

Near transfer. Performance on part 2 of the Card Rotation Task was used as the measure of near transfer, as it involved rotating two-dimensional shapes. Counter to our hypothesis, the High Variety group and the Low Variety group did not show a significant difference in their performance on the near transfer posttest, $t(185) = 0.14$, $p = 0.89$. The group that trained with High Variety did not do better on this task than the group that trained with Low Variety.

Moderate transfer. Performance on the Cube Comparison task and the Shepard and Metzler 3D Mental Rotation Task comprised our measure of moderate transfer, as they involved mentally rotating three-dimensional shapes. Counter to our hypothesis, the High Variety group and the Low Variety group did not show a significant difference in their performance on the moderate transfer posttest, $t(185) = 0.91$, $p = 0.37$. The group that trained with High Variety did not do better on these tasks than the group that trained with Low Variety.

Far transfer. Performance on the Surface Development task and the Geologic Block Cross Sectioning task comprised our measure of far transfer, as they involved spatial visualization that could benefit from improved mental rotation ability but were not necessarily measures of mental rotation (Hegarty & Waller, 2004). Counter to our hypothesis, the High Variety group and the Low Variety group did not show a significant difference in their performance on the near transfer posttest, $t(185) = 1.71$, $p = 0.89$. The group that trained with High Variety did not do better on these tasks than the group that trained with Low Variety.

Performance on Posttests: All Groups

The following supplemental analyses used the un-standardized, un-collapsed data, looking at each training group separately (Tetris shapes, Polygons, Chemistry shapes, and High Variety group), and at each posttest individually. Table 2 shows the means and standard deviations for these analyses.

Card Rotation test. ANOVA revealed no significant difference between the 4 groups on Card Rotation posttest accuracy, $F\ (3,183) = 1.39$, $p = 0.25$. An ANCOVA using

Card Rotation pretest as a covariate also revealed no significant difference among groups, F (3, 183) = 0.63, p = 0.60. No training group performed differently than any other training group on the Card Rotation test after an hour of training with their respective shape type.

Cube Comparison test. ANOVA revealed no significant difference among the 4 groups on Cube Comparison accuracy, F (3,183) = 1.18, p = 0.32. No training group performed differently than any other training group on the Cube Comparison test after an hour of training with their respective shape type.

Shepard and Metzler Mental Rotation test. ANOVA revealed no significant difference among the 4 groups on Shepard and Metzler Mental Rotation accuracy, F (3, 172) = 0.88, p = 0.45, nor on Shepard and Metzler Mental Rotation reaction time, F (3, 172) = 1.47, p = 0.23. No training group performed differently than any other training group on the Shepard and Metzler Mental Rotation test after an hour of training with their respective shape type.

Surface Development Test. ANOVA revealed no significant difference among the 4 groups on Surface Development accuracy, F (3, 183) = 1.28, p = 0.28. No training group performed differently than any other training group on the Surface Development test after an hour of training with their respective shape type.

Geologic Block Cross Sectioning Test. ANOVA revealed no significant difference among the 4 groups on Geologic Block Cross Sectioning accuracy, F (3, 183) = 0.76, p = 0.52. No training group performed differently than any other training group on the Geologic Block Cross Sectioning test after an hour of training with their respective shape type.

Discussion

The hypothesis that higher variety during training would result in worse training performance than a training task with lower variety was supported and consistent with findings from Chen and Mo (2004) and Posner and Keele (1968). This is likely because the High Variety condition reported significantly higher challenge and effort than the Low Variety condition. However, this increased challenge and effort did not lead to greater improvement in the target skill, as the High Variety Condition performed no better than the Low Variety Condition on near, moderate, or far transfer tasks. This is in contrast to findings from the literature (Gick & Holyoak, 1983; Chen & Mo, 2004; Poster & Keele, 1968), which suggests that if variety does lead to more efficacious training interventions, it is not simply because those interventions are more challenging. Variety conveys some other, less obvious benefit.

Limitations

In order to allow for a control group that trained with each shape type, this study limited its High Variety condition to just four shape categories. It is very possible that this was not enough variety to convey benefit to the trainee. For example, in the Lumosity brain-training study by Bainbridge and Mayer (2017), each group trained with 5 different game types in their respective domains yet did not improve on transfer measures. It is very likely that a much greater amount of variety would be necessary in order to buoy transfer.

The training intervention described above lasted for only an hour. While we had reason to believe that mental rotation ability could meaningfully improve after such a short intervention due to findings by Adams, Stull and Hegarty (2014), in the Lumosity brain-training study by Bainbridge and Mayer (2017) participants failed to improve after 20 hours of training. It is likely that there is a dose-response effect with training, and an hour was not enough to reveal a benefit for variety even if there was one.

The standardized composite effect size between the High Variety and Low Variety conditions was $d = 0.20$. A post hoc power analysis was conducted using this effect size. It established that we would need a total sample size of 620 participants to find a statistically significant difference between these conditions if there were one. The total sample size for this study was N = 187, which is insufficient given the effect size found.

Due to the already daunting task of recruiting participants for four conditions, a fifth, inactive control condition was excluded from this design. The inclusion of an inactive control group could have revealed if the training was simply ineffective, or if trained groups performed better than an inactive control group regardless of variety level.

Future Directions

The proposed mechanism in this intervention for why higher variety would result in greater transfer was that participants would not invest in an explicit, specific shape encoding strategy, as seen in Tarr and Pinker (1989) and would instead invest in an implicit spatial visualization strategy. However, participant strategy was only inferred by the amount of challenge and effort the participants reported. Future studies should include more direct measures of participant strategy as a way of gauging whether a variety intervention is successfully inducing a spatial visualization strategy or not.

Future studies should attempt similar interventions with longer training periods and greater variety. It is difficult to design a study with proper controls when the experimental variable is variety, but it may be necessary to sacrifice a degree of control in order to properly investigate this question. Longer training periods could demonstrate a dose-response effect in transfer performance, which makes claims of causation stronger.

Another benefit to a longer training period is it would also be less problematic to include redundant pretests and posttests. A practice effect on a measure would be so strong if administered twice in the same day that it could obscure any effect of training. If the training stretches over a few weeks, then practice effects would be less overpowering. Being able to make direct comparisons between pretest and posttest performance makes it much less problematic to make firm claims about causation and improvement in training studies.

Chapter III

Experiment II: The Variety Hypothesis in Working Memory Training

Objective

This study continues the aims of the previous study, but instead focuses on improving working memory. Training working memory has been a goal of cognitive scientists for half a century (Klingberg, 2010) because of its proposed link to general fluid intelligence (Conway, Kane & Engle, 2003; Au, et. al., 2015; Jaeggi et.al., 2010). Working memory and intelligence share a common capacity constraint: the number of items that can be held in mind and manipulated through attentional control processes. It is therefore reasonable to assume that increasing this number through training could increase one's ability to perform tasks that require fluid intelligence (Halford, Cowan, & Andrews, 2007; Jaeggi, Buschkuel, Jonides, & Perrig, 2008). This underlying relationship also makes fluid intelligence a good far transfer measure, as the surface demands of a fluid intelligence task are very dissimilar to a working memory task, yet both can improve if the number of items that can be held in mind increases.

For the purposes of this study, we use the definition of working memory put forth by Engle (2002): Working memory consists of temporary memory storage with an element of executive control that regulates and manipulates the active portion of memory. Engle's definition makes a distinction between short-term storage and working memory. He posits that the former is a component of the latter, but the latter also includes an element of attentional control; i.e. one must be able to keep items from short-term storage in mind while also performing some other attention-demanding task (Engle, 2002; Unsworth & Engle, 2007).

The n-back task is the most commonly used means of operationalizing working memory training (Au, et. al., 2015; Jaeggi et.al., 2010; Klingberg, 2010; Soveri, Antfolk, Karlsson, Salo & Laine, 2017). It involves presenting subjects with a series of stimuli and then asking them to indicate whether the current stimulus matches the one presented "n" stimuli previously. A subject performing a 1-back would need to indicate if the current stimulus was the same as the one they just saw. A subject performing a 2-back would need to indicate if the current stimulus was the same as the one before the one they just saw. The "n" gradually increases as the task goes on, requiring the subject to keep more and more items in mind as they play. Dual and single versions of the task exist, but there is no difference between them in terms of training efficacy (Buschkuehl & Jaeggi, 2010).

Literature Review

Working Memory Training. Numerous prior studies have shown that working memory and fluid intelligence can be improved with n-back training (Klingberg, 2010; Klingberg, Forssberg, & Westerberg, 2002; Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Au, et. al., 2015). An archetypical example can be seen in a study by Jaeggi, Buschkuel, Jonides, and Perrig (2008). In it they trained participants with an adaptive n-back training task. The participants were randomly assigned to either an inactive control group, or to train for 8 days, 12 days, 17 days, or 19 days. All participants were given multiple measures of fluid intelligence before and after training. All four training groups improved comparably, and gains on the training task began to level off after the fifth day of training. When the active groups were compared to the inactive control group they found a significant difference in improvement from pretest to posttest on measures of fluid intelligence with an impressive effect size of d = 0.65.

Despite this impressive finding, there is controversy among researchers about whether and how far n-back training can transfer. A meta-analysis by Soveri and colleagues (2017) found that the majority of improvements reported in n-back training studies are task-specific and do not extend beyond novel versions of the trained task. This is in contrast to the meta-analysis by Au et. al. (2015), which found that n-back training reliably improved performance on fluid intelligence measures with an, albeit modest, effect size of $g = 0.24$. The inconsistency in results between studies leaves the matter of whether n-back training is effective subject to debate to this day.

Growth mindset. Due to the potential link between n-back training and fluid intelligence, growth mindset was added as a moderator measure. Growth mindset is the belief that intelligence is malleable and can be improved. It is defined in contrast to a fixed mindset, which is the belief that intelligence is fixed and cannot be changed (Dweck, 2015). Having a growth mindset has been shown to improve various learning outcomes, and it has been suggested that the reason for this is that people with a growth mindset are more likely to keep working hard when challenged than people with a fixed mindset (Dweck, 2015). If a participant with a fixed mindset were to suspect that the purpose of the n-back training was to train intelligence, but their beliefs about intelligence suggest that intelligence cannot be improved, they may be less motivated to exert the effort necessary to improve on the n-back and may therefore benefit less from the training than someone with a growth mindset.

Theory and Predictions

In keeping with the general hypotheses of this dissertation, it is possible that the discrepancy between results in the field of working memory training could be the result of differences in variety during training. As applied to n-back training, the variety hypothesis would suggest that during the learning phase a subject who practices an n-back with limited stimuli would get very good at keeping those specific stimuli in mind, but would be unable to transfer the strategies learned and skills acquired during that training to novel situations. For example, if a participant only practiced the n-back using letters, perhaps the strategy they developed would be most applicable to letters, such as repeating the letters in the phonological loop, or relying on the familiarity of the letter shapes to notice when a similar shape reappeared. The strategy would be an *explicit* strategy. They would be consciously aware of how and why they were improving. In contrast, a participant who practiced the n-back using dissimilar and varied stimuli could not rely on an explicit strategy. Rehearsing stimuli in the phonological loop would not be as effective if the stimuli consist of complex images of landscapes, for example. Relying on familiarity to bolster recognition would not be very effective when asked to remember irregular polygons. In this way, the subject who practices with many types of stimuli cannot develop an explicit strategy, and must instead invest in a more *implicit* strategy. They would need to invest in the underlying skill of keeping a greater number of things in mind and would be find it difficult to verbalize how or why they were improving. This would, as in the schema induction literature (Chen & Mo, 2004; Posner & Keele, 1968), impede their performance during the training task compared to their low-variety counterparts, but would benefit them more when asked to apply their improvements to a novel situation.

To test whether increasing variety improves transfer after an n-back working memory training intervention, we compared participants who trained on an n-back task with just one stimulus type (letters) to participants who trained with 6 types of stimuli (letters, animals, faces, words, irregular polygons, and landscapes). It was hypothesized that:

1. The Low Variety group will perform with greater accuracy during the n-back training than the High Variety group.

2. The High Variety group will report more challenge, more effort, and more enjoyment during the training task than the Low Variety group.
3. The High Variety group will perform better on a near transfer task (an n-back using unfamiliar stimuli), moderate transfer measures of short term memory (digit span) and working memory (symmetry span, Corsi blocks), as well as far transfer measures of fluid intelligence (Raven's Advanced Progressive Matrices).
4. Self-Efficacy will moderate the effectiveness of the training.
5. Participants with a fixed mindset who perceive the n-back training to be highly related to intelligence will not invest as much in the training and will have lower accuracy on the n-back compared to participants with a growth mindset and those who do not perceive the task to be related to intelligence

Method

All sessions were conducted at a laboratory in the Department of Psychological and Brain Sciences at University of California, Santa Barbara. The room in which all sessions were conducted contained 4 Dell computers arranged in 4 separate cubicles. Up to four participants at any stage of their training could be in the lab at a time. All pretests, training, and posttests were conducted on the Dell computers using either e-prime or Direct RT software. Consent forms and questionnaires were administered on paper.

Participants

The participants consisted of 73 students from the University of California, Santa Barbara with an average age of 19.6 years (*SD* = 1.62). An additional 14 participants started the study but dropped out before the final session. They were recruited from the UCSB psychology subject pool. Forty of them were women, 29 were men, and 4 declined to state their gender. Participants were randomly assigned to one of three conditions: 27 to the low variety n-back training group, 20 to the high variety n-back training group, and 26 to the inactive control group. Control participants attended 2 sessions, totaling about 45 minutes, and were compensated $15. Active participants from both groups attended 6 sessions, totaling about 6 hours, and were compensated $60.

Materials

Low variety n-back training. The n-back task requires participants to indicate if the present item matches the one they saw *n* items prior. For example, if a participant in a 2-back task were shown the letters A + E + B + E + T + Y + T they would indicate that the 4th and 7th letters were targets. As the training progresses, the *n* gradually increases. The low-variety condition consisted of only letter stimuli. During each visit to the lab, participants completed 36 blocks. Each block consisted of 30 trials with a subset of 8 letters. Every block rotated which 8 letters were used, with 24 letters being used in total (excluding the letters V and F). If more letters were used within the same block, it would become too easy to guess if a presented letter was a match. Participants were asked to respond to every trial, pressing "1" if the stimulus matched the one presented *n* trials prior, and pressing "0" if the stimulus did not match. The ratio of *match* to *no match* trials was approximately 1:3. Each trial presented the stimulus for 4 seconds, during which time the participant could respond at any time. If the participant took longer than 4 seconds to respond, the stimulus moved on automatically and counted the trial as incorrect. After each 30 trial block the participant was given the opportunity for a brief rest. The next block started when the participant indicated they were ready. Participants received right/wrong feedback for the first block at the start of each session but were not given feedback for the remainder of the training. This was to avoid interrupting their phonological loop. Each session the participant completed 1,080 trials, and by the end of the training each participant had completed 5,400 trials.

High variety n-back training. Participants in the high variety group received a very similar training to the low variety group, with the exception that 6 categories of stimuli were used: letters, animals, faces, words, irregular polygon shapes, and landscape paintings. Examples of each of these stimuli categories can be seen in Appendix B. As in the low-variety condition, each block consisted a subset of 8 stimuli from the same category, with 24 different stimuli available in each category. After each 30-trial block the category would rotate (i.e. 30 trials of letters, then 30 trials of animals, then 30 trials of faces, etc…); participants saw each category 6 times over the course of each session, with a different subset of stimuli in each block. Aside from this rotation, every other aspect of the training remained the same as in the low-variety group, including interval length, number of trials, feedback, and breaks. Figure 9 contains a sample from each stimulus category.

Video break. To avoid cognitive fatigue, participants received numerous breaks. After each 30 trial block participants were given a short break. Halfway through each session, after 18 blocks, participants were given a longer break wherein they watched a 3-5 segment from the nature documentary *Planet Earth.* It has been shown previously that viewing nature can have a restorative effect on cognitive function (Berman, Jonidas & Kaplan, 2008).

Digit Span. The participant is asked to remember a series of digits presented visually. After each successful trial the number of digits is increased by one. The quantity of digits they can correctly recall in order served as a pre-test measure of prior short-term storage capacity, a component of working memory that can inform the researchers of the participant's prior ability before training. It was repeated as a post-test to assess improvement (Mueller & Piper, 2014).

Symmetry Span. The participant is presented with a grid of black and white squares. They must indicate whether the array is symmetrical along a vertical center line. They are then presented with another grid with one square highlighted and are instructed to remember the position of the highlighted square. They then repeat the symmetry judgment with a different array, then are presented with another highlighted square to remember. After multiple cycles, they are asked to recall all of the highlighted squares they were asked to remember by clicking a blank grid. The number of squares they must keep in mind before recall gradually increases as they successfully complete each round. The accuracy with which they recall the sequence of highlighted squares serves as the dependent variable (Mueller & Piper, 2014). Complex span tasks like the symmetry span have been shown to be reliably related to working memory capacity (Redick, et.al., 2012). A screenshot of the symmetry portion of the task can be seen in Figure 10.

Corsi Blocks. The participants are presented with 10 blocks arranged in an array. The researcher points to a series of blocks in a random order. The participant must reproduce the sequence (Klingberg, Forssberg & Westerberg, 2002). The number of blocks they can indicate in the correct order serves as the dependent variable. This is used as a posttest measure of working memory. A screenshot of the task can be seen in Figure 11.

Novel n-back task. To assess near transfer participants were asked to complete a version of the n-back task very similar to their training, but with a stimulus category neither group practiced with: flowers. The task contains 6 blocks of 30 trials each. The *n* increases every block, starting with a 1-back and ending with a 6-back. Each block contains a subset of 8 flowers out of a total of 24 flowers. Examples of the stimuli used can be seen in Figure 12. Accuracy on the 5-back and 6-back blocks served as the dependent variable.

Raven's Standard Progressive Matrices. Participants are shown a matrix of figures with one slot empty. They were asked to infer what the figure in the empty slot must be by interpreting the pattern of shapes in the other rows and columns (Jaeggi, et. al., 2011). Their score consists of how many matrices they successfully solved within 10 minutes. This was used as a posttest far transfer measure of fluid intelligence. An example trial can be seen in Figure 13.

Questionnaire. At the end of their final session all participants were given a questionnaire. Active participants received a questionnaire that consisted of 30 questions. Control participants received a subset of those questions numbering 16. The full questionnaire can be viewed in Appendix B.

Control Questionnaire. Three of these questions gathered demographics: gender, age, and race. Thirteen of the questions assessed growth mindset using binary items. An example of a growth mindset question would be "You got a bad grade on a test. Why do you think that happened? A) the subject was too difficult for me. B) I didn't study hard enough". Statements supporting a growth mindset were scored as 1 and statements supporting a fixed mindset were scored as 0, then scores were summed for a possible total score ranging from 0 to 13, with higher scores indicating higher growth mindset tendencies.

Active Questionnaire. In addition to the above questions, active participants from both training groups received questions assessing enjoyment of the training, challenge during the training, self-efficacy as related to the training task, subjective improvement over the course of the training, and if they perceived the training task as being related to intelligence or not. These questions all used a 5 point Likert scale. The full questionnaire can be viewed in Appendix B.

Procedure

All participants came to the lab and completed the digit span and the symmetry span pretests. After the pretests, participants were randomly assigned to the Low Variety, High Variety, or Control groups. Control participants were dismissed after the pretests and asked to return to the lab in 10-14 days to compete the posttests. Participants in both active groups then completed 18 blocks of the 1-back task and 18 blocks of the 2-back. During the second session active participants completed 36 blocks of the 3-back. During the third session active participants completed 36 blocks of the 4-back. During the fourth session participants completed 36 blocks of the 5-back. During the fifth session participants completed 36 blocks of the 6-back. Each session included a video break half way through the training. Each training session, including video break, took between 30 and 60 minutes to complete. During the final session both active and control participants completed the digit span, symmetry span, Corsi blocks task, novel n-back task, and Raven's Matrices task. After the posttests all participants were given the questionnaire, then were debriefed and dismissed. This final session lasted approximately 50 minutes. The study had IRB approval and was conducted in adherence with guidelines for treatment of human subjects.

Results

Results were analyzed using SPSS. Performance on the two pretest measures was standardized and averaged to create a unified measure of prior ability, which was used as a covariate on ANCOVA tests of posttest performance. Table 3 shows the means and standard deviations for all conditions on all measures.

Pretest Differences

The groups were equivalent on demographic factors. An ANOVA of age, $F(2,64) = 0.47$, $p = 0.63$, and a Chi Square analysis of gender, $X^2(2, N = 70) = 2.25$, $p = 0.33$, revealed no significant differences among groups.

The groups were also not significantly different on pretest performance. An ANOVA of digit span pretest, $F(2,69) = 1.45$, $p = 0.24$, and symmetry span pretest, $F(2,63) = 0.31$, $p = 0.73$, revealed no significant differences among groups. When performance on these pretests was standardized and combined into a measure of prior ability the trend held. An ANOVA comparing groups on this new prior ability measure revealed no significant differences, $F(2,70) = 0.99$, $p = 0.38$. This indicates that random assignment succeeded at making the groups sufficiently equivalent and any differences found between groups on posttest measures can be reasonably attributed to the training intervention.

Performance on Training Task

Performance on the training task was defined as proportion of accurate responses to whether a presented stimulus matched one presented *n* trials prior. A t-test of overall performance on all n-back training tasks indicated that the Low Variety group ($M = 0.80$, $SD = 0.05$) performed significantly better than the High Variety group ($M = 0.72$, $SD = 0.08$) on the training overall, $t(45) = 4.00$, $p < 0.00$. The results supported the hypothesis that the Low Variety group would outperform the High Variety group during the n-back training.

Table 4 shows means and standard deviations for each session of the training task, as well as t-test results for each session. It shows that the Low Variety group responds with significantly greater accuracy than the High Variety group in all but the 1-back, and that the High Variety group at times performs around chance (~66.67%).

Questionnaire Responses

Means and standard deviations for all Questionnaire measures for each group can be seen in Table 3.

Enjoyment and challenge. T-tests revealed that the High Variety group and Low Variety group did not significantly differ on measures of enjoyment, $t(43) = 0.76$, $p = 0.42$, or challenge, $t(43) = 0.32$, $p = 0.75$. There was also no correlational relationship between reported enjoyment and training performance. $r = 0.01$, $p = 0.46$, or between reported challenge and training performance, $r = 0.17$, $p = 0.13$.

Subjective improvement. A t-test revealed no significant difference between the High Variety group and the Low Variety group on subjective improvement, $t(43) = 0.65$, $p = 0.52$. There was no correlational relationship between subjective improvement and training accuracy during the final 6-back training session, $r= 0.09$, $p = 0.28$. In addition, subjective improvement did not correlate with improvement on the symmetry span, $r= 0.09$, $p = 0.31$. Nor did it significantly correlate with digit span improvement, $r = -0.24$, $p = 0.06$. Subjective improvement was significantly correlated with performance on the novel n-back posttest [$r= 0.27$, $p = 0.04$] with a medium effect size, but overall it appears to be a poor predictor of actual improvement.

Intelligence perception and growth mindset. The High Variety, Low Variety, and control groups did not differ in growth mindset, $F(2,67) = 0.14$, $p = 0.87$. The High Variety and Low Variety groups did not differ in terms of how much they thought the n-back training task related to intelligence, $t(43) = 1.63$, $p = 0.11$. It was hypothesized that participants with a fixed mindset who perceive the n-back training as related to intelligence might not invest as much in the training task, as they would have no expectation of improvement. A linear regression looking at the influence of IQ perception and growth mindset found that these factors did not significantly predict overall training accuracy. The regression was not significant, $F(2, 42) = 0.75$, $p = 0.48$, and had an r^2 of 0.03. When the factors were converted into categorical variables and analyzed using a 2(fixed vs growth mindset) x 2 (high vs low IQ perception) ANOVA with overall training accuracy as a dependent variable the ANOVA confirmed that there was no significant interaction between growth mindset and intelligence perception on training accuracy, $F(1, 42) = 0.03$, $p = 0.87$. Altogether the results do not support hypothesis 5; participants with a fixed mindset did not do more poorly during the training if they perceived the task to be highly related to intelligence.

Self-efficacy. A t-test revealed that participants in the High Variety group reported significantly lower task-specific self-efficacy than the Low Variety group, $t(43) = 2.99$, $p < 0.01$. Participants in the High Variety group did not feel as capable as the Low Variety group of doing well and improving on the n-back training. A 2x2 ANOVA comparing training condition (High vs Low Variety) and self-efficacy level (High vs Low) on overall training accuracy found no significant effect of self-efficacy on n-back training performance, $F(1,41) = 1.60$, $p = 0.21$, and no significant interaction between group and self-efficacy on n-back training performance, $F(2,41) = 0.37$, $p = 0.55$, indicating that if there was a difference in self-efficacy between groups it did not moderate the efficacy of their training.

Performance on Posttests

Analyses on posttests were conducted using ANOVA comparing group (High Variety vs Low Variety vs Control) on performance, as well as ANCOVA using either pretest performance or the averaged and standardized prior ability score as a covariate. Least

Square Difference (LSD) Post Hoc tests are for ANOVA results. Table 3 contains means and standard deviations for all groups on all posttests.

Digit Span. ANOVA revealed a significant difference between groups on digit span posttest, $F(2,68) = 3.46$, $p = 0.04$. Post Hoc LSD test showed the High Variety group (M = 7.11, SD = 1.29) performed significantly more poorly than the Control group (M = 8.08, SD = 1.26), but not significantly different from the Low Variety group (M = 7.85, SD = 1.23). An ANCOVA using digit span pretest performance as a covariate showed that pretest performance was a significant predictor of posttest performance, $F(2,66) = 25.03$, $p < 0.00$; however, using pretest performance as a covariate also made the differences between groups no longer significant, $F(2,66) = 2.64$, $p = 0.08$. When pretest performance is subtracted from posttest performance to create a digit span improvement score, the groups do not differ significantly on this improvement score according to an ANOVA, $F(2,67) = 1.45$, $p = 0.24$. The hypothesis that the High Variety group would perform better on the short term memory transfer task of digit span was not supported.

Symmetry Span. ANOVA revealed no significant differences among groups on the symmetry span posttest, $F(2,66) = 0.24$, $p = 0.78$. An ANCOVA using symmetry span pretest performance as a covariate showed that pretest performance was a significant predictor of posttest performance, $F(2,59) = 18.72$, $p < 0.00$. The groups remained not significantly different on symmetry span posttest when symmetry span pretest was used as a covariate, $F(2,59) = 0.12$, $p = 0.89$. When pretest performance is subtracted from posttest performance to create a digit span improvement score, the groups did not differ significantly on this improvement score according to an ANOVA, $F(2,60) = 0.10$, $p = 0.90$. The hypothesis that the High Variety group would perform better on the working memory transfer task of symmetry span was not supported; neither active group performed signifiantly better than the inactive control group.

Corsi Blocks. An ANOVA revealed a significant difference between groups on the Corsi blocks posttest, $F(2,68) = 3.84$, $p = 0.03$. Post Hoc LSD test showed that the High Variety group ($M = 6.17$, $SD = 1.30$) performed significantly worse than the Low Variety group ($M = 7.26$, $SD = 1.46$), but was not significantly different from the Control group ($M = 6.92$, $SD = 1.13$). When an ANCOVA was conducted using the prior ability score as a covariate the groups remained significantly different, $F(2,67) = 3.47$, $p = 0.04$. The ANCOVA also showed that prior ability was not a good predictor of performance on Corsi blocks, $F(2,67) = 0.40$, $p = 0.53$. The hypothesis that the High Variety group would perform better on the working memory transfer task of Corsi blocks was not supported, and in fact the High Variety group had worse performance than the Low Variety group after the training.

Novel N-Back. An ANOVA revealed a significant difference among groups on the novel n-back posttest, $F(2,69) = 8.75$, $p < 0.00$. Post Hoc LSD test showed that the High Variety group ($M = 22.65$, $SD = 2.39$) and the Low Variety group ($M = 22.84$, $SD = 2.69$) both performed better than the Control group ($M = 19.24$, $SD = 4.57$) but not significantly different from each other. When an ANCOVA was conducted using the prior ability score as a covariate the groups remained significantly different, $F(2,68) = 8.62$, $p < 0.00$. The ANCOVA also showed that prior ability was not a good predictor of performance on the novel n-back, $F(2,68) = 0.03$, $p = 0.87$. The hypothesis that the High Variety group would perform better on the near transfer of an n-back task with unpracticed stimuli was not supported; both trained groups performed better than the inactive control group, i.e. n-back training benefitted performance on an n-back task, but no benefit was found for increased variety during training.

Ravens Advanced Progressive Matrices. ANOVA revealed no significant differences among groups on the Raven's Matrices posttest, $F(2,70) = 0.56$, $p = 0.58$. An ANCOVA using prior ability as a covariate showed that performance on the pretests was a significant predictor of performance on Raven's Matrices, $F(2,69) = 12.08$, $p < 0.01$, supporting the notion that working memory is related to fluid intelligence. The groups remained not significantly different on Raven's Matrices posttest when prior ability was used as a covariate, $F(2,69) = 0.56$, $p = 0.58$. The hypothesis that the High Variety group would perform better on the far transfer fluid intelligence task of Raven's Matrices was not supported; neither active group performed any better than the inactive control group.

Discussion

The only hypothesis that was supported was the first one: the Low Variety group did perform better during the n-back training task than the High Variety group. We found no evidence to support the hypothesis that the High Variety group would transfer their learning to other working memory tasks better than the Low Variety group. In fact, on the digit span posttest the High Variety group performed worse than the inactive control group, and on Corsi blocks the High Variety group did worse than the Low Variety group. The only measure on which the High Variety group did significantly better than another group was the near-transfer novel n-back posttest, and even then they did no better than the Low Variety group. In addition, the High Variety group did not report different levels of enjoyment, challenge, or subjective improvement on the training task when compared to the Low Variety group.

The failure of the High Variety group to outperform the Low Variety group (and in some cases to do significantly worse than the Low Variety or control groups) could potentially be attributed to the lower self-efficacy they reported in regards to the transfer task. Perhaps the High Variety group was aware of how poorly they were performing on the n-back training task, decreasing their motivation to invest in the training and negatively affecting their learning. This account is supported by evidence that the High Variety group was performing around chance (66%) on the training task during the 2-back, 5-back, and 6-back sessions, which suggests the task may have been too difficult for them to do successfully. This account is undermined by the finding that the High Variety group did not report greater challenge than the Low Variety group. It is also undermined by the finding that subjective improvement was not significantly related with task performance or improvement measures, indicating that the participants did not have the metacognitive skills necessary to accurately gauge their performance. Most damningly, self-efficacy was not significantly related to training performance, so even if the training did induce lower self-efficacy in the High Variety group, it did not appear to affect their behavior.

Limitations

The design of this study did not include trial-by-trial feedback, as we worried it would interfere with potential strategies. However classic learning research has shown that feedback improves learning curves (Trowbridge & Cason, 1932). It is possible we impeded our participants' ability to improve by not providing adequate feedback on their performance.

Due to a combination of random assignment and drop-outs, the High Variety group (n = 20) ended up smaller than our Low Variety (n =27) and inactive control groups (n = 26). An a priori power analysis indicated that with the effect size of d = 0.65 established in the Jaeggi and colleagues (2008) study of similar design we would need a sample size of 30 per group, which this study fell just short of with an average of 24 participants per group.

The effect size established by Jaeggi and colleagues (2008) was established by comparing a group that trained with the n-back working memory training and an active control group that played Tetris. One would expect a smaller effect size when comparing two groups who both trained with the n-back task. When the High Variety and Low Variety groups were compared on standardized performance averaged across all posttests we found an effect size of $d = 0.61$ (Low Variety: $M = 0.19$, $SD = 0.58$; High Variety: $M = -0.16$, $SD = 0.57$). A post hoc power analysis was conducted using this effect size. The variety hypothesis was one-tailed, as in we expected increased variety to improve performance on transfer measures relative to lower variety. If this were the case, we would have needed a sample size of 34 per group; however, the means trended in the opposite direction, favoring the Low Variety group on most posttest measures. This indicates we should have used a two-tailed hypothesis, in which case we would have needed a sample size of 44 people per group to reveal a difference between our groups. We fell short of this sample size by almost half. If the length of the data collection had doubled, perhaps we could have shown a reliable advantage for Low Variety n-back training relative to High Variety n-back training on transfer measures.

Future Directions

A potentially fruitful new direction would be to take a mixed-methods approach to n-back training studies and use qualitative techniques to assess participant strategy. By interviewing participants about how they approach the training task we may develop a better understanding of which strategies lead to better performance during training, which strategies lead to better transfer to novel domains like fluid intelligence, and if there is a difference between strategies for these two outcomes.

Chapter IV

Experiment III: The Variety Hypothesis in Action Video Game Training

Objective

The prior two studies lacked an element of external validity. This dissertation seeks to explain why action video games are effective at training cognitive skills when brain-training games are not, and yet the prior two experiments were only somewhat game-like. While they included game-like elements, such as a computer interface and gradually increased difficulty, they lacked many of the motivational rewards and sensory stimulation of a video game, as well as the sheer amount of variety a video game affords. These simplified designs helped to limit confounds, but by including so many experimental controls the interventions may have moved too far away from the real world. The current study seeks to explore the same proposed role of variety in a more ecologically valid way by using an off-the-shelf action video game. The aim of the current study is to test whether greater variety, in the form of more diverse environments and goals, increases cognitive gains after training with an off-the-shelf action video game.

Literature Review

Multiple sources confirm the effectiveness of action video games in improving performance on measures of visual attention. Exhaustive reviews of the phenomenon have been conducted by Bediou and colleagues (2018) and Mayer (2014). In general, intervention studies comparing action video game training to a control group on measures of visual attention have an effect size between $d = 0.45$ (Bediou et al, 2018) and $d = 1.18$ (Mayer, 2014).

The current study most closely resembles an intervention study conducted by Green and Bavelier (2003). In it they took video game novices, gave them three measures of visual attention (including an Attentional Blink and Useful Field of View test), then randomly assigned them to play either *Tetris* (a puzzle game) or *Medal of Honor* (an action video game) for one hour per day for 10 consecutive days. They hypothesized that the *Medal of Honor* group would improve more on the measures than the *Tetris* group because in *Medal of Honor* attention must be distributed among many different targets in a large field of play, whereas in *Tetris* participants need only focus on one object at a time in a limited field of play. Put another way, the action video game group had to apply their visual attention skill more *variably* than the control group. Indeed, they found that participants who played *Medal of Honor* improved on all three measures compared to the *Tetris* players.

Theory and Predictions

This dissertation explores the notion that variety, the amount of different ways one is asked to apply a skill during practice, is an important factor in how action video games improve visual attention skills despite not being designed to do so. In an action video game the environment constantly changes, as do the goals. Players must apply their visual attention skill throughout, but almost never in exactly the same way or context twice. As a result, the player should find it less advantageous to develop explicit strategies, such as the conscious goal to “find the high ground”, and more advantageous to improve their more implicit skill of rapidly and accurately respond to visual stimuli in any environment in which they find themselves. If this is the case, a group that plays an action video game in a limited environment, e.g. a small area that never changes with consistent goals, should be more likely to develop explicit strategies that are specific to that environment. For example, they may find a hiding place with good cover that they can use repeatedly. These explicit strategies would improve their performance in the game environment but would not transfer to a novel environment. That hiding place with good cover may reduce the number of times they die in the game, but it won't help their performance on a transfer measure of multiple object tracking. In contrast, a group that plays an action video game in a more varied environment, e.g. an ever-changing map and goals that shift from day to day, should be less likely to invest in those explicit strategies. A hiding place may work in one map, but that knowledge is useless the next time they play. They would be better off investing in more implicit strategies like improving the automaticity with which they respond to visual stimuli, as these improvements would serve them well in any map in which they play. The same flexibility of application that make these improvements beneficial in a highly varied video game would also lead to better transfer to unrelated tests of visual attention.

To test whether greater variety improves transfer of visual attention ability after 9.5 hours of action video game play we compared novice video game players who played *Call of Duty: Black Ops* multiplayer in just one small environment to novice video game players who played *Call of Duty: Black Ops* multiplayer in a different environment every session, and then compared both of these groups to an inactive control group that played no games for two weeks. It was hypothesized that:

1. The Low Variety group will perform better (as measured by kill/death ratio) during the game training than the High Variety group.
2. The High Variety group will improve more from pretest to posttest on three measures of visual attention skill (Attentional Blink, Multiple Object Tracking, and Useful Field of View) than the Low Variety group.
3. Both of the active groups will improve more than the inactive control group.
4. The High Variety group would attribute more of their improvement to implicit, non-conscious strategies as compared to the Low Variety group, and the Low Variety

group would attribute more of their improvement to explicit, conscious strategies as compared to the High Variety group.

Method

Participants

Participants were recruited among the student population at the University of California, Santa Barbara. Of those recruited, 69 finished all 11 sessions of the study and 13 dropped out before the final session. Of these participants who finished, 23 were in the Low Variety group, 25 were in the High Variety group, and 21 were in the inactive control group. The average age was 19.63 years (*SD* = 2.10). Of those who completed all sessions, 51 were female, 17 were male, and 1 participant declined to state their gender. Participants were compensated $10 per session they attended.

Materials

All pretests, posttests, and game training was conducted in a laboratory at the University of California, Santa Barbara. The lab contains 4 desktop Dell computers equipped with Steam, e-prime, and Direct RT in four separate cubicles. Up to four participants could be in the lab at any one time. Questionnaires were administered with paper and pen.

Call of Duty: Black Ops. This game was chosen for the game training intervention because it is a first-person shooter game that shares a lot of characteristics with other action video games that have been used in effective training interventions such as *Medal of Honor, Unreal Tournament,* and other installments of the *Call of Duty* series. It has all of the elements identified by Bediou and colleagues (2018) as characteristic of games that have been shown to improve visual attention skills with routine play: it has a fast pace, a high degree of perceptual and motor load, the necessity to switch between focused attention and distributed attention, and a high degree of clutter and distraction. This particular version of *Call of Duty* was chosen because it offers a multiplayer mode that allows you to play against computer operated opponents. In most multiplayer modes you play against real players, which poses a problem for research as it is not feasible for the experimenter to control the difficulty of the opponents. In the *Call of Duty: Black Ops* multiplayer mode the researchers were able to gradually increase the difficulty of the opponents starting with "Recruit" in sessions 1-3, then upgrading to "Regular" in sessions 4-6, then to "Hardened" in sessions 7-9, and finally to the hardest level of "Veteran" during the final session. Both conditions followed the same difficulty progression. The objective in multiplayer mode is to kill as many opponents as possible while dying as infrequently as possible within the allotted time. During every session each participant's "kills" and "deaths" were recorded and a "kill/death ratio" (KDR) was calculated from these numbers as a measure of in-game performance. A higher KDR indicates that the player killed more opponents and died fewer times than a player with a lower KDR. If a KDR is below 1 it means the player died more times than they killed. A KDR above 1 indicates the opposite. It is a common measure of performance among habitual video game players that compensates for the differences between risk-averse and high-risk strategies. If high kills were the dependent variable players with a reckless strategy of running into the fray without regard to any defensive strategies would appear to be more skilled. If low deaths were the dependent variable then players who hid from all opponents and didn't engage in combat would appear to be more skilled. By making a ratio of kills:deaths players with a balance of higher kills and lower deaths are measured most favorably.

Low Variety Condition. Players assigned to the Low Variety condition played every session in the same "map", or location. The location called "Nuketown" was chosen because it was the smallest map available. The same number of opponents (6) and game mode (Free-For-All) was used for every session. The only element that changed in this condition was the difficulty of the computer opponents.

High Variety Condition. Players assigned to the High Variety condition changed maps every session. The maps chosen were the largest available. The names of the maps, in order, were “Crisis”, “Grid”, “Hanoi”, “Havana”, “Launch”, “Summit”, “Array”, “Cracked”, “Jungle”, and “WMD”. The High Variety condition alternated between “Free-For-All” mode and “Team Death Match” mode each session. A table showing what each active group did for each session can be seen in Appendix C.

Video Game Experience Survey (VGES). An abridged version of the Video Game Experience Survey developed by Terlecki and Newcombe (2005) was used to pre-screen participants. Sample items include “Have you ever played video games” and “How often (approximately) do you play video games”. Responses were converted to numerical values and summed. Potential scores ranged from 0 to 17. Only those who scored below a 9 were asked to participate. If the participant listed action video games or first person shooters as one of their favorite genres in question 10 the exclusionary threshold was lowered to 7. The mean score was 1.87, the median was 1 and the mode was 1. Appendix C contains the full survey.

Attentional Blink (AB). In this test participants are told to look for the four corners of a square and indicate after every string which corners they saw. They gaze at a fixation point, and then a string of letters is rapidly presented. Interspersed within the letters are 2 images depicting one or more corners of a square. There are either 0, 1, 3, or 8 letters in between these target stimuli. The attentional blink phenomenon makes it very difficult to perceive a second target stimulus presented between 0 and 270 milliseconds after a first target stimulus, which in the case of this task applies to trials where the target stimuli have only 0 or 1 letters between them. There are 160 trials total, and 60 of these have 0-1 letters between the target stimuli. The dependent variable is the accuracy with which the participant responds to these more difficult trials (Raymond, Shapiro, & Arnell, 1992). The test was administered using e-prime. An illustration of an attentional blink trial can be seen in Figure 14.

Multiple Object Tracking (MOT). In this test participants are shown an array of 8 identical objects. A subset of this array, ranging from 1 to 4, is briefly highlighted to distinguish them from the rest and the participant is asked to track the highlighted objects. The highlighting is removed, then all 8 objects begin to move around the screen (Pylyshyn & Storm, 1988). Once the objects stop, the participant must identify the ones that were highlighted at the start of the trial. There are 6 trials total. The dependent variable is how many objects they correctly identified on trials with 4 targets, of which there were 2. The maximum score is 8, the minimum is 0. The test was administered using DirectRT. A screenshot from the MOT can be seen in Figure 15.

Useful Field of View (UFOV). This test requires participants to focus on a fixation cross in the center of the screen. A target shape then appears along one of multiple axes, and the participant is asked to identify which axis it appeared upon. Halfway through the task the difficulty increases by adding distractor items and shortening the length of time the target appears. There are 240 trials total and the last 120 of these include the distractor items and shorter interval. The dependent variable is the accuracy with which participants respond to these more difficult trials (Edwards et. al., 2005). The test was administered using Direct RT. A screenshot from the UFOV can be seen in Figure 16.

Strategy Questionnaire. After the posttests active participants were given a questionnaire. It contained 2 items assessing subjective improvement on a Likert scale from 1 to 5 with 1 being "I got worse" and 5 being "I got a lot better". An example of a subjective improvement item is "Do you feel as though you got better at playing *Call of Duty* over the course of this experiment?". The questionnaire also contained 3 items assessing how much they attributed their improvement to explicit strategies on a Likert scale ranging from 1 to 5. An example of an explicit item would be "Did you learn any tricks that helped you perform better?" with 1 being "I didn't learn any tricks" and 5 being "all of my improvement was due to tricks I learned". Three of the items on the questionnaire assessed how much they attributed their improvement to implicit strategies on a Likert scale ranging from 1 to 5. An example of an implicit item would be "Did you feel as though your gameplay got more automatic?" with 1 being "No, I improved for other reasons" and 5 being "All of my improvement was due to increased automaticity". At the end of the survey the participants were asked to list the tricks and deliberate strategies they learned. Their responses were coded so that each unique strategy received a point and the points were summed to create a "strategy list" score. It was surmised that participants who relied more on environment-specific explicit strategies would be able to list more items in this section than participants who relied more on implicit improvements in their visual attention ability, so they should receive a higher "strategy list" score. The full survey is included in Appendix C.

Procedure

Before the experiment began potential participants were asked to complete the VGES and return it to the researcher. The survey was then scored and participants who received below the exclusionary threshold were invited to participate. Participants were randomly assigned to the Low Variety, High Variety, or inactive control group. The control group attended two sessions spaced 2 weeks apart. The active groups attended 11 sessions over the course of 14 days. During the first session all participants took the AB, the MOT, and the UFOV, which took 20-30 minutes. After these pretests the control group was asked to leave and the active group filled the remainder of the hour playing *Call of Duty: Black Ops* multiplayer in their assigned condition. During sessions 2 through 10 active participants played in their assigned condition for the full hour, resulting in a cumulative 9.5 hours of game play within a two week time frame. A schedule of their gameplay can be seen in Appendix C. In the final session the control participants returned and all participants re-took the AB, MOT, and UFOV. After the posttests control participants were thanked for their time, paid, and debriefed. Active participants took the strategy questionnaire and were then thanked, paid, and debriefed. This final session took about 30 minutes and always took place at least a day after the last game training session for active participants. The study had IRB approval and was conducted in adherence with guidelines for treatment of human subjects.

Results

Statistics were computed using SPSS. Means and standard deviations for all groups on all measures can be found in Table 5. Session-by-session game performance means and standard deviations can be found in Table 6.

Pretest Differences

Analysis of Variance (ANOVA) revealed no significant differences among groups on age, $F(2,64) = 0.87$, $p = 0.43$, or on prior video game experience, $F(2,66) = 0.40$, $p = 0.67$. There was also no significant difference between groups on pretest performance on the AB, $F(2,64) = 0.31$, $p = 0.74$; MOT, $F(2,64) = 0.88$, $p = 0.42$; or the UFOV, $F(2,51) = 0.52$, $p = 0.60$. This indicates that random assignment succeeded in making the groups roughly equivalent, and any differences in posttest performance among these groups can be attributed to the game training intervention.

Performance on Training Task

Performance while playing *Call of Duty: Black Ops,* as measured by KDR, differed significantly between the groups. A t-test revealed that the Low Variety group performed significantly better during the training than the High Variety group, $t(46) = -3.02$, $p < 0.01$. This supports the first hypothesis. Table 6 contains the means and standard deviations for KDR in each hour-long session of game play.

Role of Prior Experience

Participants with a large amount of prior video game experience were excluded from the study, but amount of prior video game experience still varied somewhat between participants and could have influenced the efficacy of the intervention. To test whether the limited amount of experience our participants had could have influenced their performance, pretest results were standardized and averaged across the three measures, then this score was correlated with prior video game experience scores. The correlation was significant, $r = 0.32$, $p < 0.01$, indicating that prior video game experience was a good predictor of performance on measures of visual attention. However, when prior video game experience scores were used as a covariate in ANCOVAs of posttest performance by group it did not change any of the results and was not a significant predictor of performance on any of the three visual attention posttest measures

When prior experience is correlated with training performance an explanation emerges for the discrepancy between the significant influence of prior experience on pretests and its lack of influence on posttests. While the correlation of prior experience and KDR on the early sessions of game training is trending towards significance, $r = 0.23$, $p = 0.06$, the correlation of prior experience and KDR on the last session of game training is not significant, $r = 0.06$, $p = 0.97$. This indicates that the amount of training the participants received was enough to overcome the influence of prior experience among our sample of novice video game players by the end of the game intervention.

Performance on Posttests

For each posttest two analyses were run. The first result is an ANCOVA comparing all groups (High Variety, Low Variety, and inactive control) on posttest performance using pretest performance as a covariate. In the second analysis the two active groups (High Variety and Low Variety) were collapsed into one group and compared to the inactive control group using an ANCOVA of posttest performance using pretest performance as a covariate. If a significant result was found a post hoc LSD test was conducted using improvement (posttest performance minus pretest performance) as the dependent variable. Table 5 contains means and standard deviations for all measures.

Attentional Blink (AB). An ANCOVA comparing groups on AB posttest performance using AB pretest performance as a covariate revealed no significant differences, $F(2,63) = 1.65$, $p = 0.20$, indicating that there was no difference between the High Variety, Low Variety, or control groups on AB accuracy after the training period.

When the High Variety and Low Variety groups were combined and the ANCOVA was run comparing active participants ($M = 0.53$, $SD = 0.27$) to control participants ($M = 0.51$, $SD = 0.34$) on AB posttest performance using AB pretest performance as a covariate there was still no significant difference, $F(2,64) = 3.24$, $p = 0.08$. This indicates that the game training did not sufficiently improve performance on AB accuracy more than can be attributed to practice effects. Neither hypothesis 2 nor 3 were supported with AB performance.

Multiple Object Tracking (MOT). An ANCOVA comparing groups on MOT posttest performance using MOT pretest performance as a covariate revealed no significant differences, $F(2,63) = 0.04$, $p = 0.96$, indicating that there was no difference between the High Variety, Low Variety, or control groups on the number of objects that could be successfully tracked after the training period.

When the High Variety and Low Variety groups were combined and the ANCOVA was run comparing active participants ($M = 4.33$, $SD = 1.52$) to control participants ($M = 4.48$, $SD = 1.12$) on MOT posttest performance using MOT pretest performance as a covariate there was no significant difference, $F(2,64) = 0.00$, $p = 0.99$. This indicates that the game training did not sufficiently improve performance the ability to successfully track multiple objects beyond what can be attributed to practice effects. Neither hypothesis 2 nor 3 were supported with MOT performance.

Useful Field of View (UFOV). An ANCOVA comparing groups on UFOV posttest performance using UFOV pretest performance as a covariate revealed a significant difference between groups, $F(2,45) = 3.657$, $p = 0.03$. A LSD post hoc analysis of improvement scores suggests that the Low Variety group ($M = 0.14$, $SD = 0.09$) improved more than the High Variety group ($M = 0.07$, $SD = 0.08$), but neither improved significantly more than the control group ($M = 0.10$, $SD = 0.07$). This directly opposes hypothesis 2.

When the High Variety and Low Variety groups were combined and the ANCOVA was run comparing active participants ($M = 0.56$, $SD = 0.20$) to inactive participants ($M = 0.55$, $SD = 0.16$) the significant difference evaporates, $F(2,46) = 0.00$, $p = 1.00$. Game play did not improve participants' ability to identify target shapes during the most difficult UFOV trials. Neither hypothesis 2 nor 3 were supported with UFOV performance.

Questionnaire Responses

Subjective Improvement. A t-test revealed no difference between the High Variety group (*M* = 3.53, *SD* = 0.69) and the Low Variety group (*M* = 3.65, *SD* = 0.57) on reported subjective improvement, $t(46) = -0.67$, p =0.51. There was no correlation between subjective improvement and actual improvement on KDR from the start to end of the game training intervention, $r = 0.05$, $p = 0.37$,

Subjective improvement was not correlated with actual improvement on the AB, r = 0.02, $p = 0.46$. There was also no correlation between subjective improvement and actual improvement on the MOT, $r = 0.21$, $p = 0.09$, or the UFOV, $r = 0.24$, $p = 0.08$. A participant reporting that they felt as though they had improved was not a good predictor of whether they had actually improved from pretest to posttest on any measure.

Strategy. It was hypothesized that the High Variety group would report more implicit strategies and the Low Variety group would report more implicit strategies. A t-test was used to see if the High Variety group differed the Low Variety group on implicit strategy items. It revealed no significant differences, $t(46) = 0.30$, $p = 0.77$. This indicates that the High Variety group was not more likely to attribute their improvement to implicit strategies than the Low Variety group. Similarly, a t-test comparing the High Variety group to the Low Variety group on explicit strategy items revealed no significant differences, $t(46) = -1.22$, p = 0.23. This indicates that the Low Variety group was not more likely to attribute their improvement to explicit strategies than the High Variety group. In addition, a t-test comparing the number of strategies the groups could list revealed no significant differences between the High Variety group (*M* = 2.12, *SD* = 1.20) and Low Variety group (*M* = 2.12, *SD* = 1.20), $t(45) = -0.50$, p =0.62. The Low Variety group was not able to name a higher number of explicit strategies that they used during the game training than the High Variety group. Overall there was no evidence to support Hypothesis 4.

Discussion

We hypothesized that the Low Variety group would perform better during the game training than the High Variety group. This hypothesis was supported by the data; however, the remaining hypotheses were not. The groups did not differ significantly on two of the three tests of visual attention. While there was a significant difference on the UFOV, this was the result of the High Variety group improving significantly less than the Low Variety group, and neither group improved any more than the inactive control group. If anything, increased variety hurt performance on this measure. There were no significant differences between the High Variety and Low Variety groups on strategy; however, given that subjective improvement was not related to actual improvement, it is possible that the participants lacked the metacognition necessary to make accurate judgments about what strategies they used.

In addition to failing to find a difference between playing an action video game with high variety or low variety, we also failed to find a difference between participants who played an action video game for 9.5 hours and inactive control participants. This is in direct contradiction to prior studies that have shown that action video game interventions improve performance on measures of visual attention (Bediou et. al, 2018; Green & Bavelier, 2003; Mayer, 2014).

A potential explanation for this boundary condition may lie in how the game training intervention was operationalized. The game training in this study was conducted using *multiplayer mode* in order to manipulate the element of variety while still being able to compare performance between the groups. In multiplayer mode there is no narrative element. There are no characters and no story. The only goal is to kill as many opponents as possible. There is no clear reason for this goal aside from the presumed desire to get the highest score by the end of the game. In contrast, most video game interventions have participants play through the *campaign mode* of the game. In the campaign mode of a typical action video game you play a character and the gameplay has a narrative. Each level has a goal that is justified by the story. For example, maybe the enemies are after the top-secret information in your possession and you must not allow them to get ahold of it, or perhaps your partner is being held hostage and you must break into an enemy facility to rescue them. It is possible that these narrative elements help motivate participants to push themselves to perform at the upper limit of their ability. If you have grown attached to the character of your partner, you may be more likely to persevere through a challenging level that places greater demands on your visual attention ability than you would otherwise withstand in order to save them. Perhaps narrative, not variety, is the element that sets action video games apart from brain-training games. Maybe players are more likely to sustain the effort necessary to push the upper limit of their abilities when they are invested in a story.

Limitations

The sample size for the current study was low due to a combination of factors. It was hard to recruit people to take part in a study that required them to come to the lab every weekday and at least one weekend day. Many of the people who did express interest were deemed too experienced to take part. Even among those who did agree to take part, a sizable number (n = 13) failed to finish the study. An a priori power analysis suggested that, given the effect sizes established in prior meta-analyses (Bediou et. al., 2018, Mayer, 2014), between 20 and 124 participants would have been necessary to reveal a significant result between the active group and the inactive group. Enough participants did complete the current study to end up in that range (N = 69); however, revealing a difference between the two active groups would likely require even more participants. A post hoc power analysis using the effect sizes observed in this study on improvement between the High Variety and Low Variety groups (d = 0.24) suggested a sample size of at least 432 participants would have been necessary to reveal a significant difference if one existed.

It is also possible that true video game novices are much rarer now than they were in 2003, when the Green and Bavelier experiment was published. It is possible that casual exposure to video games spaced throughout childhood had already conveyed at least some benefit to the participants' visual attention skills despite their self-identification as non-video game players. If this is the case, it will be less and less likely that similar results will be found again as video games increase in ubiquity.

Through conversations with a handful of participants after the completion of the study I discovered that some players did not enjoy the game. One player said she disliked the violent goal of multiplayer mode and spent as many of her sessions as possible in a remote corner avoiding conflict with the computer opponents. This lack of enjoyment and subsequent withdrawal from gameplay may be mitigated by the motivating narrative elements of the campaign mode of the game.

Future Directions

Future studies should explore the role of narrative in action video game training. By comparing a group that plays through campaign mode to one that plays for an equivalent amount of time in the narrative-free multiplayer mode, researchers can establish if narrative is an important factor in motivating participants to perform at the upper limit of their ability and create genuine improvements. If this is the case, a follow-up study could attempt to add narrative elements to brain-training games to increase the likelihood that players improve on the target domains.

Chapter V

General Conclusion

Summary

Overall across three experiments the variety hypothesis was not supported. In none of the studies included here did the High Variety group outperform the Low Variety group on any measure, no matter how long the training intervention or how great the amount of variety. In some cases, such as the digit span and Corsi block posttests from Experiment II or the UFOV posttest from Experiment III, the High Variety group actually performed *worse* relative to other groups. This suggests that adding greater variety to training interventions is at the very least not an effective way to improve transfer; in fact, additional variety during learning may *hurt* the likelihood of transfer. Granted, the power for the studies in this collection was lower than would be needed to show a statistically significant difference with the respective effect sizes of each study, but given that the means for each condition favor the Low Variety group rather than the High Variety group, there is no reason to believe that adequate power would have actually changed this conclusion. If anything, a larger number of participants would have more likely shown an advantage for the Low Variety training conditions.

Theoretical Implications

This result contradicts the findings from previous research (Posner & Keele, 1968; Chen & Mo, 2004, Gick & Holyoak, 1983; Soloman & Perkins, 1989). One possibility for this is that the schemas developed in the Posner and Keele (1968) experiment and the Chen and Mo (2004) experiments were not implicit, but explicit. Perhaps the participants in those studies were consciously aware of the schemas they were developing and were able to mindfully direct the development of those schemas. If this were the case, these studies would be more similar to the Gick and Holyoak (1983) experiments, in which the participants were transferring explicit schemas from one story to the next. This opens the possibility that variety does *not* aid low road transfer (which is more implicit) but *does* aid high road transfer (which is more explicit). It is likely that transfer can take numerous forms, even beyond the dichotomous low-road and high-road, and that not all instances of transfer occur via the same mechanism. Perhaps variety helps in some instances but not in others. This dissertation focused on low-road transfer of cognitive skills, and the conclusion that variety is not helpful for transfer can only be applied to this narrow category of transfer.

Empirical Contributions

These studies meaningfully contribute to the cognitive skill training literature. In particular, Experiment II failed to provide evidence that n-back working memory training improved anything other than performance on an unpracticed n-back task. This supports Soveri and colleagues' (2017) assertions that n-back training does not meaningfully improve anything other than performance on an n-back task.

Interestingly, Experiment III not only failed to find an effect of variety, it also failed to find an effect of video game training regardless of variety condition compared to an inactive control. This contradicts numerous experiments and meta-analyses that have found action video game training to be effective at improving cognitive domains such as visual attention. While some researchers have cast doubt on this trend (Unsworth, Redick, McMillan, Hambrick, Kane, & Engle, 2015; Irons, Remington, & Mclean, 2011) the bulk of the evidence suggests that action video games are indeed effective at improving performance on measures of related skills. Experiment III represents a boundary condition and suggests that this effect is not reliable under all circumstances. The confounds inherent in an intervention as complex as an action video game are numerous. The games differ greatly between titles and platforms, and there are individual player differences that could influence outcomes as well. Parsing how, why, and when video games result in cognitive improvements represents a potentially never-ending stream of research. The results of Experiment III suggest a closer look needs to be taken at the factors that differed between the Green and Bavelier (2003) version of the intervention and the current one to establish why one was effective and the other was not. Perhaps our population of novices was more experienced than the population used by Green and Bavelier (2003). Another possibility is the difference between an active and inactive control group. It may also be the case that the participants in Experiment III did not enjoy the intervention as much as the participants in Green and Bavelier's (2003) intervention. The most likely culprit is that Green and Bavelier (2003) used the campaign mode of the game and Experiment III used multiplayer mode. This final possibility will be discussed more below.

Practical Implications

The results of these studies strongly suggest that increasing variety in surface details and environments is not a useful addition to brain-training game design. In fact, this addition will at the very least hurt players' ability to perform well within the game and may even lead to decreased self-efficacy and weaker transfer.

The current studies also suggest that action video games are not a completely reliable way of improving visual attention skill in players. If one hopes to improve their visual attention skill, the greater body of research suggests playing action video games is a good way to do so, but evidently there are nuances, such as what game should be selected, how often and for how long it should be played, and whether the game includes a narrative, that should be considered before training is attempted. Playing *Call of Duty* in multiplayer mode for 9.5 hours over 14 days is apparently not sufficient to accomplish this goal.

Future Directions

A future study should directly compare a group that plays *Call of Duty: Black Ops* in multiplayer mode (no narrative) to a group that plays *Call of Duty: Black Ops* in campaign mode (with narrative). If the presence of a narrative increases motivation and encourages players to work harder than they otherwise would, the group that plays in campaign mode should see gains similar to those seen in Green and Bavelier's (2003) intervention, and the group that plays in multiplayer mode should see no gains over an inactive control group, as seen in Experiment III. Results demonstrating that this is the case would have specific and actionable implications for brain-training game design and could perhaps lead to more effective training interventions in the future. It may be the case that traditional gamification (such as rewards and leveling-up) as it is usually applied to brain-training games falls short of the full motivational power of a video game with a story.

While the variety hypothesis was not supported when it came to low-road transfer and implicit practice, these studies leave open the possibility that high-road transfer and explicit practice could still benefit from the addition of variety, such as when Gick and Holyoak (1983) successfully boosted the solution rates in Dunker's radiation problem by adding additional examples of the convergence solution. A future study could compare learners' ability to transfer knowledge to a novel situation after a lesson taught in one way with just one type of example or a lesson taught for a comparable amount of time but with many examples. An intervention like this could occur both with traditional lessons and with video games.

This collection of studies only manipulated variety along one dimension: dissimilar surface features; however, variety can be operationalized in many ways. In Experiment I we could have operationalized variety in terms of degree of rotation. The Low Variety group could have rotated shapes up to 20 degrees, whereas the High Variety group could have rotated shapes up to 160 degrees. In Experiment II variety could have been operationalized by the *n* in *n -back*. The Low Variety group could have practiced with letter stimuli up to a 3-back, whereas the High Variety group could have practiced with letter stimuli up to a 6-back or higher. Both of these examples introduce level of difficulty as a confound, but other designs that would add variety to structural features rather than surface details are possible and should be explored.

During my postdoctoral research at University of Colorado, Boulder I will engage in research that can further both of these lines of inquiry. I will be comparing learning from a traditional physics lesson to learning from a physics lesson supplemented by a physics puzzle game called *Physics Playground* (Shute, Ventura & Kim, 2013). One of the first steps is manipulating motivational supports to the game in an attempt boost player engagement and effort. Over the course of this appointment, I will have the opportunity to explore whether adding additional puzzle examples to the game and/or adding additional traditional classroom examples can boost high-road transfer of student physics knowledge. If this succeeds, we can conclude that variety is helpful for explicit learning but not implicit learning. If it fails, we can conclude that additional variety in learning does not benefit transfer.

References

"A Consensus on the Brain Training Industry from the Scientific Community," Max Planck Institute for Human Development and Stanford Center on Longevity, accessed 7/30/19, http://longevity.stanford.edu/a-consensus-on-the-brain-training-industry-from-the-scientific-community-2/

Adams, D. M., Stull, A. T., & Hegarty, M. (2014). Effects of mental and manual rotation training on mental and manual rotation performance. *Spatial Cognition & Computation*, *14*(3), 169-198.

Allaire, J., Bäckman, L., Balota, D., Bavelier, D., Bjork, R., & Bower, G. (2014). A consensus on the brain training industry from the scientific community. *Max Planck institute for human development and stanford center on longevity*.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic Bulletin & Review, 22*(2), 366-377.

Bainbridge, K., & Mayer, R. E. (2018). Shining the light of research on Lumosity. *Journal of Cognitive Enhancement, 2*(1), 43-62.

Barnett, S.M., & Ceci, S.J. (2002) When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612-637.

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin, 144*(1), 77.

Bransford, J.D., & Schwartz, D.L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24,* 61-100.

Buschkuehl, M., & Jaeggi, S. M. (2010). Improving intelligence: A literature review. *Swiss Medical Weekly, 140*(1920).

Chen, Z., & Mo, L. (2004). Schema induction in problem solving: A multidimensional analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(3), 583.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends In Cognitive Sciences, 7*(12), 547-552.

Cooper, L. A. (1975). Mental rotation of random two-dimensional shapes. *Cognitive Psychology, 7*(1), 20-43.

Cooper, L.A., & Sweller (1987). Effect of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79*(4), 347.

Day, S.B, & Goldstone, R.L. (2012). The import of knowledge export: connecting findings and theories of transfer of learning. *Educational Psychologist, 47*(3), 153-176.

Dweck, C. (2015). Carol Dweck revisits the growth mindset. *Education Week*, *35*(5), 20-24.

Edwards, J. D., Vance, D. E., Wadley, V. G., Cissell, G. M., Roenker, D. L., & Ball, K. K. (2005). Reliability and validity of useful field of view test scores as administered by personal computer. *Journal Of Clinical And Experimental Neuropsychology, 27*(5), 529-543.

Ekstrom, R. B., Dermen, D., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests* (Vol. 102): Educational testing service Princeton, NJ.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions In Psychological Science, 11*(1), 19-23.

Gee, J. P. (2007). *Good video games+ good learning: Collected essays on video games, learning, and literacy*: Peter Lang.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive psychology, 15*(1), 1-38.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*(6939), 534.

Green, C. S., & Bavelier, D. (2015). Action video game training for cognitive enhancement. *Current Opinion in Behavioral Sciences, 4*, 103-108.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in cognitive sciences, 11*(6), 236-242.

Hardy, J. L., Drescher, D., Sarkar, K., Kellett, G., & Scanlon, M. (2011). Enhancing visual attention and working memory with a web-based cognitive training program. *Mensa Research Journal, 42*(2), 13-20.

Hegarty, M., & Waller, D. (2005). Individual differences in spatial abilities. *The Cambridge handbook of visuospatial thinking*, 121-169.

Irons, J. L., Remington, R. W., & McLean, J. P. (2011). Not so fast: Rethinking the effects of action video games on attentional capacity. *Australian Journal of Psychology, 63*(4), 224-231.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, 105*(19), 6829-6833.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short-and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences, 108*(25), 10081-10086.

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—implications for training and transfer. *Intelligence, 38*(6), 625-635.

Kable, J. W., Caulfield, M. K., Falcone, M., McConnell, M., Bernardo, L., Parthasarathi, T., . . . Hornik, R. (2017). No effect of commercial cognitive training on brain activity, choice behavior, or cognitive performance. *Journal of Neuroscience, 37*(31), 7390-7402.

Klingberg, T. (2010). Training and plasticity of working memory. *Trends In Cognitive Sciences, 14*(7), 317-324.

Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal Of Clinical And Experimental Neuropsychology, 24*(6), 781-791.

Lorant-Royer, S., Spiess, V., Goncalves, J., & Lieury, A. (2008). Programmes d'entraînement cérébral et performances cognitives: efficacité, motivation... ou «marketing»? De la Gym-Cerveau au programme du Dr Kawashima. *Bulletin De Psychologie* (6), 531-549.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs, 54*(6)

Mayer, R. E. (2014). *Computer games for learning: An evidence-based approach*: Cambridge, MA: MIT Press.

McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world*: New York: Penguin.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental psychology, 49*(2), 270.

Newcombe, N. S. (2010). Picture this: Increasing math and science learning by improving spatial thinking. *American Educator, 34*(2), 29.

Ormand, C. J., Manduca, C., Shipley, T. F., Tikoff, B., Harwood, C. L., Atit, K., & Boone, A. P. (2014). Evaluating geoscience students' spatial thinking skills in a multi-institutional classroom study. *Journal of Geoscience Education, 62*(1), 146-154.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology, 77*(3p1), 353.

Prensky, M. (2006). *Don't bother me, Mom, I'm learning! How computer and video games are preparing your kids for 21st century success and how.* St. Paul: Paragon House, cop. 2006.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 179-197.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal Of Experimental Psychology: Human Perception And Performance, 18*(3), 849.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational Psychologist, 24*(2), 113-142.

Shaffer, D. W. (2006). *How computer games help children learn*. New York: Macmillan.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701-703.

Shepard, S., & Metzler, D. (1988). Mental rotation: effects of dimensionality of objects and type of task. *Journal Of Experimental Psychology: Human Perception And Performance*, *14*(1), 3.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & education, 80*, 58-67.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in newton's playground. *The Journal of Educational Research*, *106*(6), 423-430.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "brain-training" programs work? *Psychological Science in the Public Interest*, *17*(3), 103-186.

Sims, V. K., & Mayer, R. E. (2002). Domain specificity of spatial expertise: The case of video game players. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *16*(1), 97-115.

Singley, M. K., & Anderson, J. R. (1989). The Study of Transfer. In *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin & Review, 24*(4), 1077-1096.

Squire, K. (2011). *Video games and learning: Teaching and participatory culture in the digital age*. New York: Teacher's College Press

Stieff, M. (2011). When is a molecule three dimensional? A task-specific role for imagistic reasoning in advanced chemistry. *Science Education, 95*(2), 310-336.

Stieff, M., & Uttal, D. (2015). How much can spatial training improve STEM achievement? *Educational Psychology Review, 27*(4), 607-615.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21*(2), 233-282.

Terlecki, M. S., & Newcombe, N. S. (2005). How important is the digital divide? The relation of computer and videogame usage to gender differences in mental rotation ability. *Sex Roles, 53*(5-6), 433-441.

Terlecki, M. S., Newcombe, N. S., & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied Cognitive Psychology, 22*(7), 996-1013.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: an examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133*(6), 1038.

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.

Uttal, D. H., & Cohen, C. A. (2012). Spatial thinking and STEM education: When, why, and how? In *Psychology of learning and motivation* (Vol. 57, pp. 147-181): Elsevier.

Uttal, D. H., & Meadow, N. G. (2013) The Psychology of Practice: Lessons From Spatial Cognition. In *The Oxford Handbook of Cognitive Psychology*. (pp. 874-885): Oxford University Press.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817.

Tables & Figures

Table 1
Standardized and averaged means and standard deviations for Experiment I

Group	*Mean*	*SD*
Training Accuracy*		
Low Variety	92.81	5.99
High Variety	89.66	6.29
Training Reaction Time (in milliseconds)*		
Low Variety	2042.94	926.71
High Variety	2471.60	767.86
Challenge (Scale of 1-5)*		
Low Variety	2.06	0.97
High Variety	2.39	0.91
Effort (Scale of 1-5)*		
Low Variety	2.91	1.17
High Variety	3.33	1.05
Enjoyment (Scale of 1-5)		
Low Variety	3.10	0.92
High Variety	3.25	0.94
Composite Posttest Performance		
Low Variety	0.03	0.62
High Variety	-0.10	0.65
Near Transfer		
Low Variety	0.02	0.96
High Variety	-0.02	1.07
Medium Transfer		
Low Variety	0.03	0.76
High Variety	-0.09	0.74
Far Transfer		
Low Variety	0.06	0.89
High Variety	-0.18	0.76

Note. Asterisk (*) indicates significant difference between groups based on t-test. n = 138 for the Low Variety group and n = 49 for the High Variety group.

Table 2
Means and standard deviations for all groups in Experiment I

Group	*Mean*	*SD*
Average Training Accuracy*		
Tetris	93.42	3.66
Polygons	88.29	7.06
Chemistry	97.14	2.55
Mixed	89.66	6.29
Average Training Reaction Time*		
Tetris	1935.56	646.96
Polygons	2682.50	940.70
Chemistry	1441.86	739.68
Mixed	2471.60	767.87
Card Rotation Pre (Number correct out of 80 items)		
Tetris	56.02	16.86
Polygons	51.72	7.06
Chemistry	58.10	13.91
Mixed	55.10	18.09
Card Rotation Post (Number correct out of 80 items)		
Tetris	59.18	14.54
Polygons	57.30	10.81
Chemistry	62.98	12.88
Mixed	59.37	14.14
Cube Comparisons (Number correct out of 42 items)		
Tetris	28.70	7.42
Polygons	26.30	7.70
Chemistry	26.95	6.66
Mixed	26.45	6.65
3D Mental Rotation RT (in milliseconds)		
Tetris	3949.82	1224.85
Polygons	3923.02	1635.75
Chemistry	4472.47	1623.17
Mixed	3881.35	1375.48
3D Mental Rotation Accuracy (Number correct out of ## trials)		
Tetris	48.52	7.40
Polygons	46.18	11.18
Chemistry	49.38	10.31
Mixed	47.42	9.41

Surface Development (Number correct out of 60 items)		
Tetris	37.42	14.93
Polygons	34.89	14.23
Chemistry	36.73	14.57
Mixed	31.96	16.00
Geologic Block Cross Sectioning (Number correct out of 16 items)		
Tetris	3.94	3.24
Polygons	4.17	3.29
Chemistry	3.59	2.84
Mixed	3.33	2.38
Challenge (Scale of 1-5)		
Tetris	2.08	0.94
Polygons	2.36	0.99
Chemistry*	1.68	0.85
Mixed	2.39	0.91
Effort (Scale of 1-5)		
Tetris	3.04	1.19
Polygons	3.23	1.00
Chemistry*	2.39	0.97
Mixed	3.33	1.05
Enjoyment (Scale of 1-5)		
Tetris	3.20	0.88
Polygons	3.06	0.96
Chemistry	3.01	0.93
Mixed	3.25	0.94
Self Efficacy (Scale of 1-5)		
Tetris	4.01	0.62
Polygons	4.02	0.60
Chemistry	3.93	0.57
Mixed	3.86	0.63

Note. Asterisk (*) indicates significant difference from other groups based on ANOVA and Tukey post-hoc tests. n = 50 for the Tetris group, n = 47 for the Polygons group, n = 41 for the chemistry group, and n = 49 for the Mixed group.

Table 3
Means and standard deviations for Experiment II

Group	*Mean*	*SD*
Training Accuracy		
Low Variety	0.80	0.05
High Variety	0.72	0.08
Challenge (Scale of 1-5)		
Low Variety	3.65	0.68
High Variety	3.58	0.62
Enjoyment (Scale of 1-5)		
Low Variety	3.32	0.93
High Variety	3.07	1.25
Subjective Improvement (Scale of 1-5)		
Low Variety	3.42	0.75
High Variety	3.28	0.67
IQ Precept (Scale of 1-5)		
Low Variety	3.04	0.87
High Variety	2.64	0.70
Self-Efficacy (Scale of 1-5)*		
Low Variety	3.03	0.57
High Variety	2.54	0.47
Age		
Low Variety	19.72	1.62
High Variety	19.28	1.32
Control	19.71	1.83
Growth Mindset (Scale of 0-13)		
Low Variety	9.67	1.92
High Variety	9.61	1.09
Control	9.88	1.83
Digit Span Pre		
Low Variety	7.56	1.58
High Variety	6.68	1.57
Control	7.23	1.93
Digit Span Post*		
Low Variety	7.85	1.23
High Variety	7.11	1.29
Control	8.08	1.26

Symmetry Span Pre		
Low Variety	0.35	0.25
High Variety	0.28	0.23
Control	0.32	0.26
Symmetry Span Post		
Low Variety	0.43	0.30
High Variety	0.38	0.29
Control	0.43	0.27
Corsi Blocks*		
Low Variety	7.26	1.46
High Variety	6.17	1.30
Control	6.92	1.13
Novel N-Back*		
Low Variety	22.84	2.69
High Variety	22.65	2.39
Control	19.24	4.57
Raven's Matrices		
Low Variety	11.30	3.28
High Variety	10.40	2.74
Control	10.46	3.87

Note. Asterisk (*) indicates significant difference from control group based on t-test. *n* = 27 for the Low Variety group, *n* = 20 for the High Variety group, and *n* = 26 for the Control Group

Table 4
N-Back Training Performance by Block, Experiment II

	Low Variety		High Variety		t-test result
	Mean	SD	Mean	SD	
1-back	0.95	0.10	0.87	0.17	$t(27.95) = 1.766, p = 0.09$
2-back	0.77	0.12	0.64	0.13	$t(41) = 3.51, p = 0.01$*
3-back	0.79	0.12	0.71	0.10	$t(44) = 2.68, p = 0.01$*
4-back	0.81	0.10	0.73	0.09	$t(45) = 2.81, p = 0.01$*
5-back	0.73	0.07	0.68	0.06	$t(44) = 2.56, p = 0.01$*
6-back	0.71	0.06	0.67	0.07	$t(45) = 2.34, p = 0.02$*

Table 5
Means and standard deviations for all measures in Experiment III

Group	*Mean*	*SD*
Age		
Low Variety	20.00	2.09
High Variety	19.68	2.66
Control	19.15	1.09
Prior Video Game Experience		
Low Variety	2.13	2.30
High Variety	1.60	2.04
Control	1.76	1.92
Overall Kill/Death Ratio (KDR) during game training*		
Low Variety	1.39	0.52
High Variety	0.97	0.42
Attentional Blink Pretest Accuracy		
Low Variety	0.38	0.23
High Variety	0.35	0.23
Control	0.41	0.30
Attentional Blink Posttest Accuracy		
Low Variety	0.55	0.25
High Variety	0.51	0.28
Control	0.51	0.34
Multiple Object Tracking Pretest (0-8 objects)		
Low Variety	3.83	1.72
High Variety	3.67	1.72
Control	4.29	1.55
Multiple Object Tracking Posttest (0-8 objects)		
Low Variety	4.41	1.56
High Variety	4.24	1.48
Control	4.48	1.12
Useful Field of View Pretest Accuracy		
Low Variety	0.42	0.21
High Variety	0.48	0.18
Control	0.45	0.15
Useful Field of View Posttest Accuracy*		
Low Variety	0.58	0.24
High Variety	0.53	0.19
Control	0.51	0.18

Explicit Strategies (Scale of 1-5)		
Low Variety	3.26	0.67
High Variety	3.04	0.59
Implicit Strategies (Scale of 1-5)		
Low Variety	2.99	0.76
High Variety	4.05	0.67

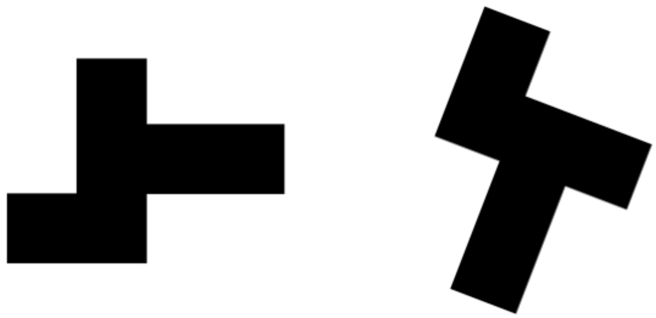
Note. Asterisk (*) indicates significant difference from control group based on t-test. *n* = 23 for the Low Variety group, *n* = 25 for the High Variety group, and *n* = 21 for the Control Group

Table 6

Action video game performance as measured by Kill:Death ratio in Experiment III.

	High Variety		Low Variety	
	Mean	SD	Mean	SD
KDR2	1.27	0.76	1.77	0.80
KDR3	1.80	1.06	2.34	1.26
KDR4	1.08	0.76	1.67	1.10
KDR5	0.92	0.38	1.43	0.72
KDR6	1.15	0.58	1.56	0.60
KDR7	0.66	0.39	0.99	0.48
KDR8	0.64	0.58	1.00	0.68
KDR9	0.61	0.34	1.01	0.45
KDR10	0.52	0.58	0.64	0.47

Figure 1

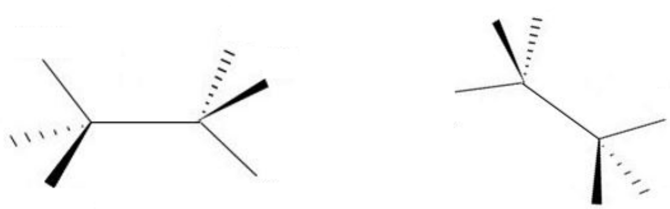


Sample Tetris shapes from mental rotation training task in Experiment I

Figure 2

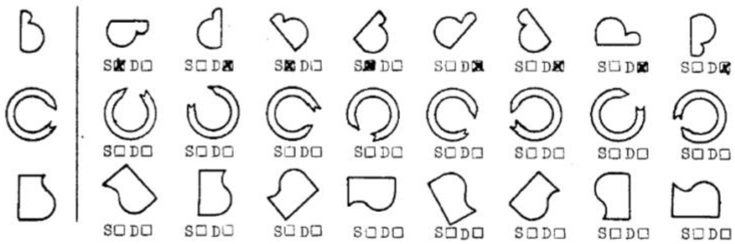
Sample Polygons from mental rotation training task in Experiment I.

Figure 3



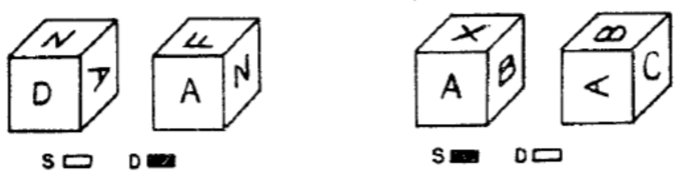
Sample Chemistry shapes from mental rotation training task in Experiment I.

Figure 4



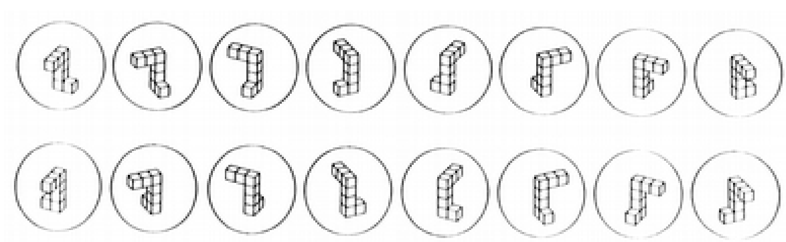
Sample items from Card Rotation test in Experiment I.

Figure 5



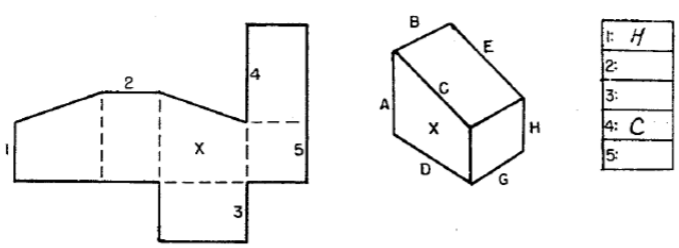
Sample items from Cube Comparison test in Experiment I.

Figure 6



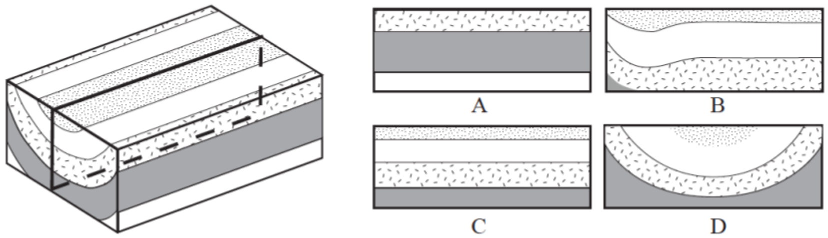
Sample items from Shepard and Metzler Mental Rotation test in Experiment I.

Figure 7



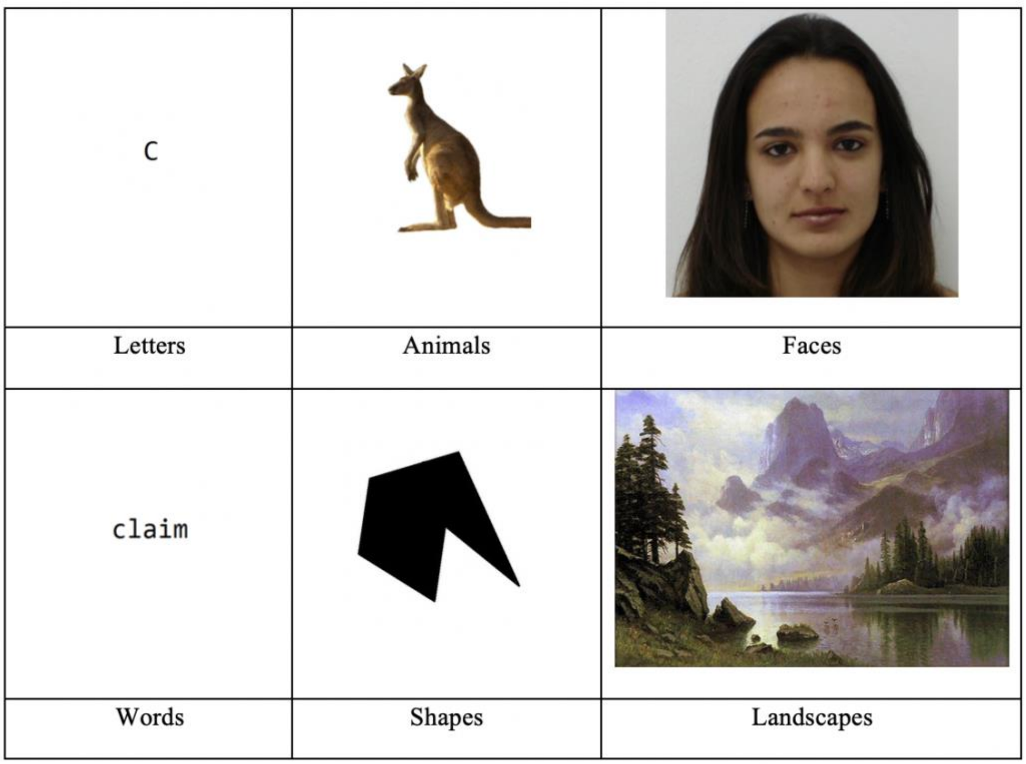
Sample item from Surface Development test in Experiment I.

Figure 8



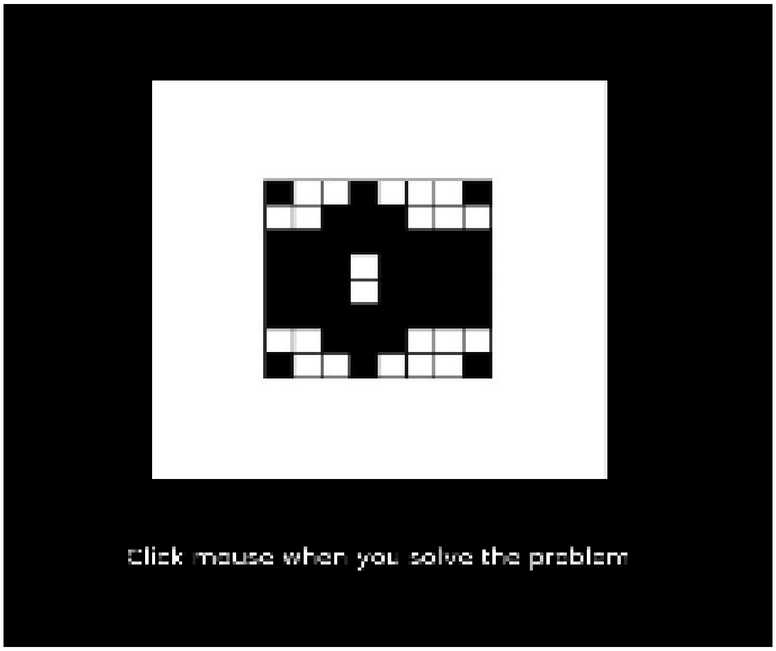
Sample item from Geologic Block Cross Sectioning test in Experiment I.

Figure 9



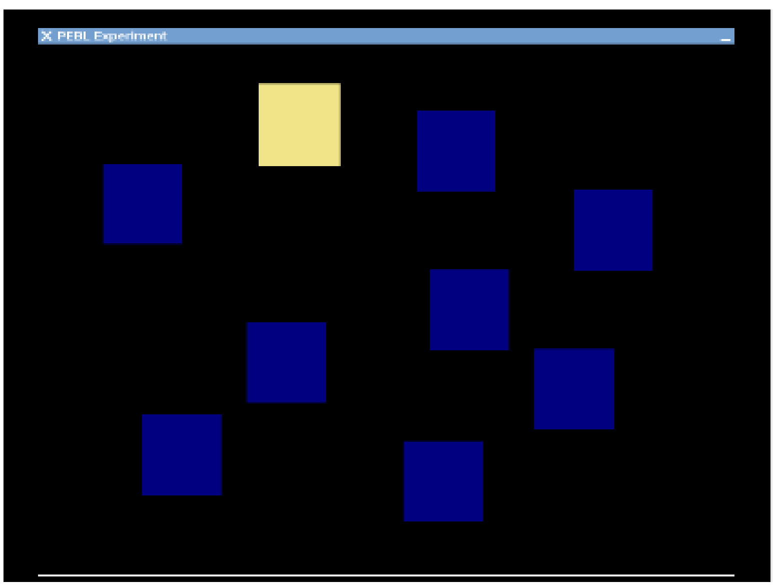
Sample stimuli from working memory training task in Experiment II.

Figure 10



Screenshot from Symmetry Span test in Experiment II.

Figure 11

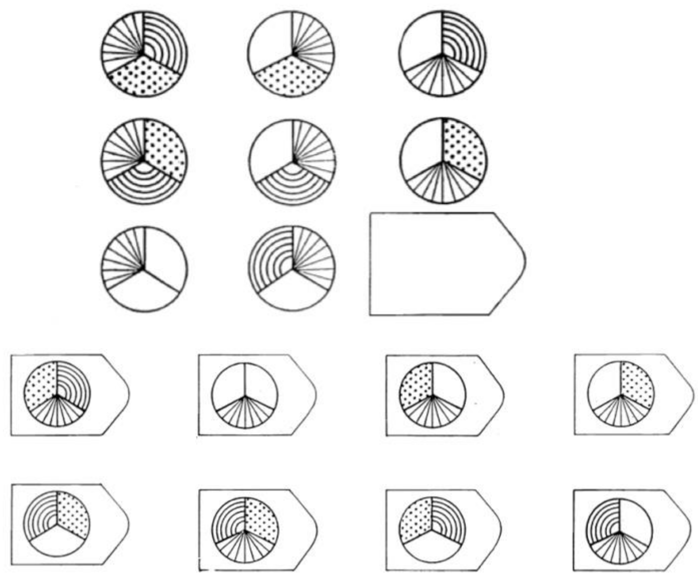


Screenshot from Corsi block test in Experiment II.

Figure 12

Examples of flower stimuli from novel n-back posttest in Experiment II.

Figure 13

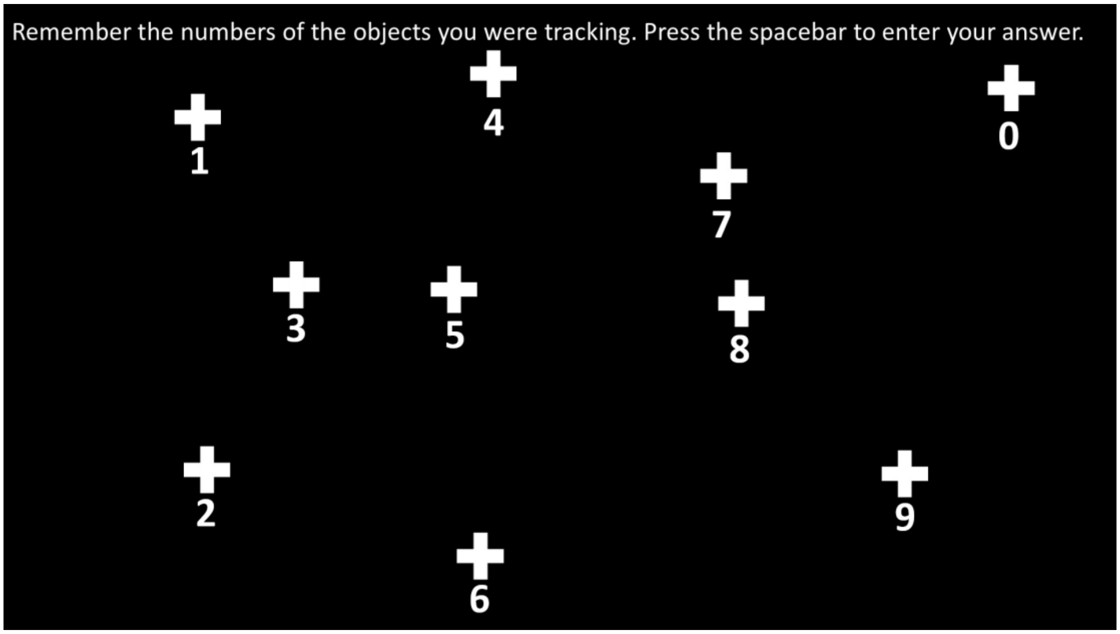


Sample trial from Raven's standard progressive matrices posttest in Experiment II.

Figure 14

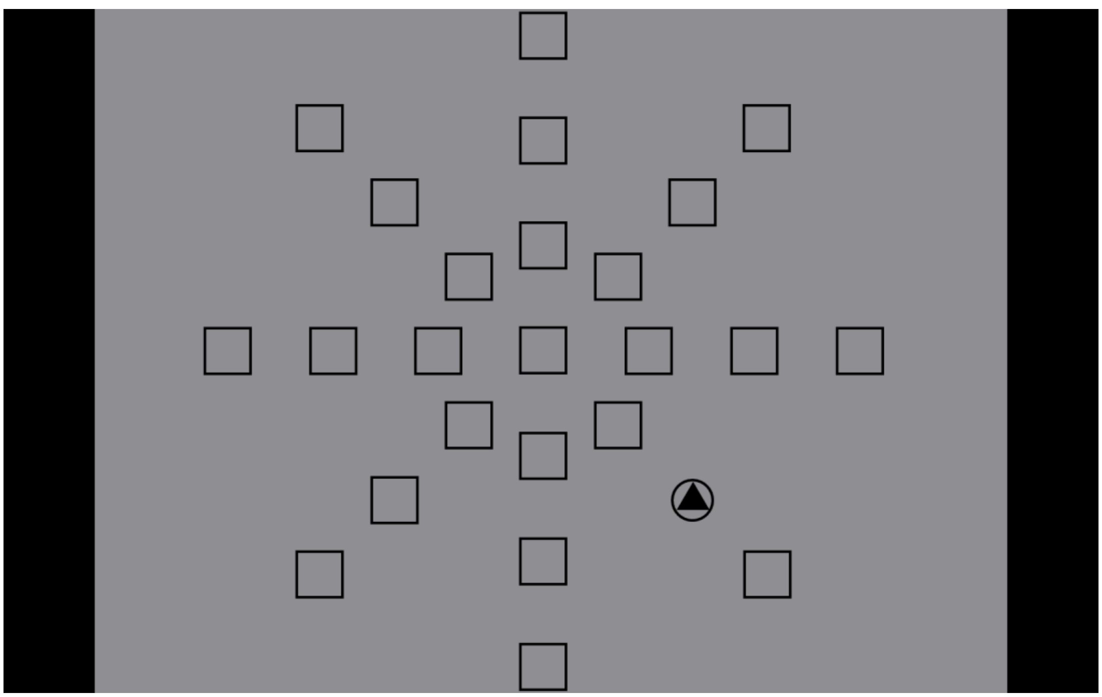
An illustration of the sequence of stimuli in one trial of the attentional blink test from Experiment III.

Figure 15



A screenshot from the multiple object tracking (MOT) test from Experiment III.

Figure 16



A screenshot from the useful field of view (UFOV) test from Experiment III.

Appendix A

Posttest Questionnaire for active conditions in Experiment I:

1. I enjoyed training with this task.

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
1	2	3	4	5

2. I would like to train with more tasks like this.

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
1	2	3	4	5

3. Please rate how appealing this training was for you.

Very unappealing	Somewhat unappealing	Medium	Somewhat appealing	Very appealing
1	2	3	4	5

4. Please rate how difficult this training was for you.

Very easy	Somewhat easy	Medium	Somewhat difficult	Very difficult
1	2	3	4	5

5. Please rate how much effort you exerted in this training task.

Very low	Somewhat low	Medium	Somewhat high	Very high
1	2	3	4	5

Appendix B

Complete posttest questionnaire from Experiment II:

1. I enjoyed training with this task.

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
1	2	3	4	5

2. I would like to train with more tasks like this.

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
1	2	3	4	5

3. Please rate how appealing this training was for you.

Very unappealing	Somewhat unappealing	Medium	Somewhat appealing	Very appealing
1	2	3	4	5

4. Please rate how difficult this training was for you.

Very easy	Somewhat easy	Medium	Somewhat difficult	Very difficult
1	2	3	4	5

5. Please rate how much effort you exerted in this training task.

Very low	Somewhat low	Medium	Somewhat high	Very high
1	2	3	4	5

Please rate how true the following statements feel to you

6. My ability to hold multiple things in mind has improved

Not at all true	Not really	Neutral	Somewhat true	Very true
1	2	3	4	5

7. I feel smarter now than I did before I started this study

Not at all true	Not really	Neutral	Somewhat true	Very true
1	2	3	4	5

8. A smart person is more likely to do well on this task than a dumb person

Not at all true	Not really	Neutral	Somewhat true	Very true
1	2	3	4	5

9. Improving on this task means you have gotten smarter

Not at all true	Not really	Neutral	Somewhat true	Very true
1	2	3	4	5

10. I got better at the training task over the course of this study

Not at all true	Not really	Neutral	Somewhat true	Very true
1	2	3	4	5

11. I felt like I was good at the training task

Not at all true	Not really	Neutral	Somewhat true	Very true
1	2	3	4	5

12. I felt like I was capable of succeeding at the training task

Not at all true	Not Really	Neutral	Somewhat true	Very true
1	2	3	4	5

13. I felt like nothing could make me better at the training task

Not at all true	Not really	Neutral	Somewhat true	Very true
1	2	3	4	5

14. I felt like there was no point in trying to get the right answers on the training task

Not at all true	Not really	Neutral	Somewhat true	Very true
1	2	3	4	5

Choose the answer that best applies to you:

15. You failed to figure out a riddle. Why do you think that happened?

A) I am not clever enough to figure it out
B) I didn’t think hard enough about it

16. You get a bad grade on a test. Why do you think that happened?

A) The subject is too difficult for me
B) I didn’t study hard enough

17. You lost at a board game. Why?

a) Luck was not on my side for this game, but it may be next time.
b) I am not as good as the other players

18. If you were to get a bad score on the SAT:

A) I would not bother retaking it because it probably wouldn’t change much
B) I would schedule another test and study harder for the next one in the hopes of improving my score

19. If you were to get a bad score on the SAT:

A) I would feel like that reflected poorly on my abilities
B) I would think I could do better under better conditions

20. Failing at something makes me want to:

A) not try it again
B) Try a different method

21. Feedback on my performance makes me feel more:

A) Unmotivated
B) Motivated

22. You had to Google how to solve a puzzle in a video game. Why?

a) I would never have figured it out on my own because I'm not good at puzzles
b) It was designed to be too difficult to solve without getting a hint

23. You turned in a draft for a class, and when you get it back it has a lot of edits. Which best describes how you would react?

a) I would put off looking at it because I'm embarrassed of the mistakes I must have made
b) I would be eager to incorporate the edits to improve the paper

24. Which is a more likely reason for why you read the chapter the teacher assigned?

A) It was interesting and I wanted to know more about it
B) I want to prove I'm a good student

25. A teacher tells you that you don't have any talent for their subject. How would you react?

a) Stop investing in the class and either resign myself to a low grade or drop the course
b) Change my approach, either by switching teachers or trying harder

26. Someone did poorly in high school. Why do you think that happened?

A) They weren't trying very hard
B) They're not a very good student

28 What is your gender?

A) Male
B) Female
C) Transgender (M to F)
D) Transgender (F to M)
E) Gender queer/non-binary
F) Other ______________

29. What is your age? ________

Appendix C

Training schedule for High Variety and Low Variety groups in Experiment III:

GROUP 1

Session	1	2	3	4	5	6	7	8	9	10
Map	Crisis	Grid	Hanoi	Havana	Launch	Summit	Array	Cracked	Jungle	WMD
Game Mode	Team Death Match	Free-for-all	Team Death Match	Free-For-All	Team Death Match	Free-For-All	Hardcore TDM	Hardcore Free-For-All	Hardcore TDM	Hardcore Free-For-All
Difficulty	Recruit	Recruit	Recruit	Regular	Regular	Regular	Hardened	Hardened	Hardened	Veteran

GROUP 2

Session	1	2	3	4	5	6	7	8	9	10
Map	Nuketown	Nuketown	Nuketown	Nuketown	Nuketown	Nuketown	Nuketown	Nuketown	Nuketown	Nuketown
Game Mode	Free-for-all	Free-for-all	Free-for-all	Free-For-All	Free-for-all	Free-For-All	Free-for-all	Free-for-all	Free-for-all	Free-for-all
Difficulty	Recruit	Recruit	Recruit	Regular	Regular	Regular	Hardened	Hardened	Hardened	Veteran

Video Game Experience Survey (VGES) used to assess prior video game experience:

	Have you ever played video games?	Yes	No
	Do you currently play video games?	Yes	No

If your answer was "no" to either question do not answer the remaining questions and return this survey to the experimenter

7. How long have you been playing Video Games?

6 months	1 year	2-5 years	5-10 years	10 or more years

8. How often (approximately) do you currently play video games?

Daily	Weekly	Once a Month	Once in 6 months	Once a year	Less than once a year or never

9. How good do you feel you are at playing video games?

Very Good	Moderately Good	Not very skilled	No skill

10. What are your Top 3 (in order) genres, or video game categories, that you enjoy to play? (Choose from the list below)

1st:
2nd:
3rd:

Video Game Genres (for #10)			
Action	Survival Horror	Flight	Educational
Fighting	God games	Racing	Maze
First-Person Shooter	Stealth	Sports	Music
Role Playing	Strategy Wargames	Military	Pinball
MMORPG	Adventure	Space	Platform
Simulators	Arcade	Strategy	Puzzle
Real-time Strategy/ Turn-based strategy	Realtime Tactical/ turn based tactical	City Building Games	Economic simulation games
Vehicular Combat	other (specify)		

Strategy questionnaire given to active participants in Experiment III:

Do you feel as though you got better at playing Call of Duty over the course of this experiment?

1	2	3	4	5
I got worse	I got no better	I got a little better	I got better	I got a lot better

Do you feel as though your visual attention skills improved over the course of playing the game?

1	2	3	4	
It got worse	It got no better	It got a little better	It got better	It got a lot better

How much of your improvement can be attributed to conscious methods that you could describe to someone else?

1	2	3	4	5
None, I improved for other reasons	A few, but they weren't responsible for much of my improvement	Some of my improvement was due to conscious decisions	Most of my improvement was due to conscious decisions	All of my improvement was due to conscious decisions

How much of your improvement can be attributed to unconscious improvements that would be hard to describe to someone else?

1	2	3	4	5
None, I improved for other reasons	A few, but they weren't responsible for much of my improvement	Some of my improvement was due to unconscious improvements	Most of my improvement was due to unconscious improvements	All of my improvement was due to unconscious improvements

How much of your improvement can be attributed to deliberate strategies that you found to be particularly effective?

1	2	3	4	5
None, I improved for other reasons	A few, but they weren't responsible for much of my improvement	Some of my improvement was due to deliberate strategies	Most of my improvement was due to deliberate strategies	All of my improvement was due to deliberate strategies

How much of your improvement can be attributed to “muscle memory”?

1	2	3	4	5
None, I improved for other reasons	A few, but they weren’t responsible for much of my improvement	Some of my improvement was due to “muscle memory”	Most of my improvement was due to “muscle memory”	All of my improvement was due to “muscle memory”

Did you learn any tricks that helped you perform better?

1	2	3	4	5
I didn’t learn any tricks	A few, but they weren’t responsible for much of my improvement	Some of my improvement was due to tricks I learned	Most of my improvement was due to tricks I learned	All of my improvement was due to tricks I learned

Did you feel as though your gameplay got more automatic?

1	2	3	4	5
None, I improved for other reasons	A little, but this wasn’t responsible for much of my improvement	Some of my improvement was due to increased automaticity	Most of my improvement was due to increased automaticity	All of my improvement was due to increased automaticity

If you did learn any tricks or deliberate strategies, what were they?