

**Researching and Developing the Impacts
of Virtual Identity on Computational Learning
Environments**

by

Dominic Kao

Submitted to the Department of Electrical Engineering
and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2018

© Massachusetts Institute of Technology, 2018. All rights reserved.

Author.....

Department of Electrical Engineering
and Computer Science
January 25, 2018

Certified by

D. Fox Harrell
Professor of Digital Media and Artificial Intelligence
Thesis Supervisor

Accepted by

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

**Researching and Developing the Impacts
of Virtual Identity on Computational Learning
Environments**

by Dominic Kao

Submitted to the Department of Electrical Engineering
and Computer Science on January 25, 2018,
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

With the current proliferation of educational games, MOOCs, and with the pervasive use of virtual identities such as avatars in systems ranging from online forums to virtual reality simulations, it is increasingly important to understand the impacts of avatars. Over two years, I led an initiative in MIT's Imagination, Computation, and Expression (ICE) Laboratory conducting experiments involving > 10,000 participants to understand the impacts of virtual identities on users in virtual environments. Using a computer science learning platform and game of our own creation as an experimental setting, we have been studying the impacts of avatar use on users' performance and engagement in computer science learning environments. This is a topic of increasing importance in human-computer interaction [69, 130, 132, 310, 452, 549]. While a great deal of work focuses on procedural thinking and problem solving, we argue that attending to learners' identities and their engagement to be equally important. We systematically explored the impacts of different avatar types on users, beginning with distinctions between anthropomorphic vs. non-anthropomorphic avatars, user likeness vs. non-likeness avatars, and other conditions informed by insights from the learning sciences and sociology. Our studies have revealed that avatars can support, or harm, performance and engagement. Several notable trends are: 1) simple abstract avatars (such as geometric shapes) are especially effective when the player is experiencing failure, e.g., while debugging, 2) likeness avatars (avatars in a user's likeness) are not always effective, 3) role model avatars (in particular scientist avatars) are often effective, and 4) successful likeness avatars that are a user's likeness when doing well and otherwise abstract are effective. We describe our studies leading to these findings and end with a follow-up study.

Thesis Supervisor: D. Fox Harrell
Title: Professor of Digital Media and Artificial Intelligence

Dedicated to my father.

Acknowledgments

To my advisor, Fox Harrell, for your wisdom, encouragement, and taking a chance on me, my thanks are beyond words. To my committee members Rob Miller, Karen Brennan, and Patrick Winston, your exemplary teaching and optimism have made indelible impressions on my life, I am so genuinely grateful. To various faculty at MIT including Nick Montfort, Scot Osterweil, Hal Abelson, Edward Schiappa, Sam Madden, Gerald Jay Sussman, Frans Kaashoek, and Michael Sipser, thank you for helping me along the way. To my labmates Sneha Veeragoudar Harrell, Chong-U Lim, Pablo Ortiz, Peter Mawhorter, Sercan Şengün, Danielle Olson, Laurel Carney, Aziria Rodríguez, Ali Jahanian, Pierre Tchetgen, James Bowie-Wilson, Maya Wagoner, Ainsley Sutherland, Ayse Gursoy, Sonny Sidhu, Jia Zhang, Erica Deahl, and Jason Lipshin, it was a pleasure to work with you. To Yael Niv, Carlos Diuk, Adam Finkelstein, Andy Bavier, Vincent Gaudet, Amir Alimohammad, and Cliff Kondratiuk, it was your support that made this possible. To researchers I've met, thank you for your collegiality. To anonymous reviewers, thank you for making me a better researcher. To my brothers, relatives, and friends in the USA, Canada, and Taiwan, thank you for being there all along. To my parents, the depth of my appreciation cannot be fully described here. To my wife, Huai-Li, thank you for everything.

Contents

Cover page	1
Abstract	3
Acknowledgments	7
Contents	9
List of Figures	13
List of Tables	17
1 Introduction	19
1.1 Motivation	19
1.2 MazeStar Platform	20
1.3 Experimental Overview	20
1.4 Design Principles	21
1.5 Findings Summary	24
1.6 Generalizability	26
2 Related Work	29
2.1 AIR Project	29
2.2 Blended Identities	30
2.3 Identification and Similarity	31
2.4 Stereotyping	32

2.4.1	Stereotype Threat	32
2.5	Avatar Impacts on Engagement and Performance	36
2.6	Constructionism	36
2.6.1	Constructionism’s Beginnings	36
2.6.2	Constructionism in the Present	37
2.6.3	Modding and Constructionism	38
2.6.4	Instantaneous Selves	39
2.7	Computational Thinking	40
2.7.1	What Is Computational Thinking?	40
2.7.2	Initiatives	43
2.7.3	Criticism	43
2.8	Adaptive Learning	44
2.9	Other Systems/Games That Teach Computer Science	45
3	The MazeStar Platform	47
3.1	The Game	47
3.1.1	Command Types	47
3.1.2	Command Limits and Compilation Errors	49
3.1.3	Level Overview	50
3.1.4	Correct Solutions	50
3.1.5	Computing Concepts Covered	53
3.1.6	Previous Versions	53
3.2	Experiment Controller	56
3.3	The Editor	57
3.3.1	Editor Basics	57
3.3.2	Stickers	58
3.3.3	Custom Images	58
3.3.4	Testing a Map	59
3.3.5	Sharing a Map	59
3.3.6	Example Player-Created Maps	60

3.3.7	Image Search	61
3.3.8	Databases	69
3.3.9	Session Handler	70
3.3.10	Avatar Creators	71
3.3.11	Computing Concepts Covered	73
4	Experimental Overview	75
4.1	Methods Overview	75
4.1.1	Measures	76
4.1.2	Recruitment	80
4.1.3	Design	81
4.1.4	Data Analyses	82
4.2	Data Overview	82
4.2.1	Participants Overview	82
4.2.2	Summative Data Overview	84
4.3	Experimental Trajectory	87
4.3.1	Trajectory	91
4.3.2	Main Findings	92
5	Experiments	95
5.1	Central Avatar Experiments	98
5.1.1	Shape vs. Likeness #1/#2	98
5.1.2	NoAvatar vs. Likeness	109
5.1.3	Shape vs. Friend	109
5.1.4	Likeness vs. EasyLikeness	111
5.1.5	ScientistText vs. ShapeText	111
5.1.6	Shape vs. Scientist	112
5.1.7	Shape vs. InstantLikeness	113
5.1.8	Shape vs. RoleModel	113
5.1.9	Shape vs. Scientist vs. Athlete	123
5.1.10	Successful Likeness	129

5.2	Additional Avatar Experiments	138
5.2.1	Phantoms vs. Non-Phantoms	138
5.2.2	Red vs. Blue	147
5.3	Interface Experiments	154
5.3.1	Feedback Positive vs. Negative vs. Neutral vs. Nothing	154
5.3.2	Mini-Game Loss vs. Near-Win vs. Win	164
5.3.3	Game Theme Basic vs. Circuit vs. RPG vs. Choice	166
5.3.4	Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice . .	183
5.4	Culminating Experiment	183
5.4.1	Badge Type Comparison	183
5.5	Chapter References	211
6	Conclusion	213
6.1	Main Findings	213
6.2	Domain Applications	214
6.3	Limitations	214
6.4	Future Work	215
6.5	Closing Reflections	216
A	Mazzy Version Listing	217
B	Calculations	219
C	Protocol Versions	221
D	Additional Tables	225
	Appendix	217
	Bibliography	245

List of Figures

1-1	MazeStar platform components.	21
1-2	Sample avatars from the successful likeness experiment.	24
2-1	Basic components of computational identity applications.	30
3-1	MazeStar platform components.	48
3-2	Level 1 in Mazzy introduces the basic game mechanics	48
3-3	Level 6 introduces looping	48
3-4	JIT hint to click the panel	54
3-5	JIT hint to use arrow keys	54
3-6	Levels in first version of Mazzy.	55
3-7	Animated tutorials.	56
3-8	Blank 11x11 map.	57
3-9	Grass and water tiles.	57
3-10	Stickers.	58
3-11	Searching for “cat”.	58
3-12	Placed search item.	59
3-13	Playing the “cat” map.	59
3-14	Sharing a direct link.	60
3-15	Sharing a website.	60
3-16	Generated webpage.	60
3-17	Map feedback form.	60
3-18	“HomeRoad” features a variety of assets. M/29.	61

3-19	“Lava Jump” has a basic, but effective design. M/20.	62
3-20	“Picnic Time” creatively uses stickers to make the path “fuzzier.” F/24.	62
3-21	“Summer Days” creatively uses tiles/stickers to create a horizon and sky. F/36.	63
3-22	“The ground is Lava” is a simple and visually cohesive map. M/31.	63
3-23	“Starcraft theme” is . . . a starcraft theme. M/29.	64
3-24	“Map1” requires going the long way. M/27.	64
3-25	“Island volcano” is a large 30x30 map. M/32.	65
3-26	“LavaZone” uses a variety of different-colored tiles to good (gradient) effect. M/25.	65
3-27	“4Corners” features different habitats. M/40.	66
3-28	“Crossing the Stream” requires a large number of staggered jumps. F/24.	66
3-29	“solar system” features different planets. M/24.	67
3-30	“Perils and party guys” is colorful. F/32.	67
3-31	“jennymap” is also colorful, and has a “retrographics look.” F/29.	68
3-32	“Garden” is a long map like many scrolling games. M/21.	68
3-33	“The Ground is Lava” features an unorthodox use of perspective. F/35.	68
3-34	“Twitter Logo” features a critique of social media. F/17.	69
3-35	MazeStar login screen.	70
3-36	Built-in avatar creator.	71
3-37	Example avatar faces.	72
3-38	Two full-body avatar examples.	72
3-39	Creating a Mii.	73
4-1	Demographic Profile	83
4-2	Example Data Slice; Upper	85
4-3	Example Data Slice; Lower	86
4-4	Performance Averages	88
4-5	Engagement Averages	89
4-6	Playtime Averages	90
5-1	Sample avatars.	99

5-2	Words used to describe likeness avatars. Larger corresponds to higher recurrence.	105
5-3	Words used to describe shape avatars. Larger corresponds to higher recurrence.	106
5-4	Experiment 2 game ratings and level completion averages between avatar types across social categories. Here, the focus is on two social groups underrepresented in STEM (95% CIs).	110
5-5	Player selected role model avatars.	114
5-6	Player selected geometric shape avatars.	115
5-7	Avatar Ratings.	119
5-8	Average Enjoyment.	120
5-9	Scientists.	125
5-10	Athletes.	125
5-11	Shapes.	125
5-12	Game Experience Questionnaire (GEQ) responses for all participants. . . .	128
5-13	Game Experience Questionnaire (GEQ) responses for female participants. .	128
5-14	Sample avatars.	133
5-15	Performance.	135
5-16	Game Experience Questionnaire (GEQ) responses for all participants. . . .	137
5-17	In the videogame <i>Dark Souls</i> two apparitions (white figures on the left and right) show the player (center) what happens when the player character runs directly at the dragon.	140
5-18	Apparitions network.	142
5-19	Avatars.	149
5-20	Game Experience Questionnaire (GEQ) responses for all participants. . . .	151
5-21	Game Experience Questionnaire (GEQ) responses for male and female participants.	151
5-22	Positive condition.	157
5-23	Negative condition.	157
5-24	Neutral condition.	158
5-25	None condition.	158

5-26	Game Experience Questionnaire (GEQ) responses.	160
5-27	Levels 1-4	174
5-28	Levels 5-8	175
5-29	Levels 9-12	176
5-30	Choice Condition	177
5-31	Performance—Graphs	178
5-32	Self-Efficacy—Graph	179
5-33	GEQ—Graph	180
5-34	PENS—Graph	181
5-35	Conditions: a) Role Model, b) Personal Interest, c) Achievement, and d) Choice.	195
5-36	Badges as they appear: a) In-game, and b) In-editor.	195
5-37	Mii avatar creator.	200
5-38	Measures, Post-Game. Error bars are standard error of the mean (SEM). . .	204
5-39	Measures, Post-Editor. Error bars show SEM.	205
5-40	Example maps rated overall 2 (left), 4 (center), and 6 (right).	210

List of Tables

1.1	Experiment Summary	23
3.1	Mazzy Commands	50
3.2	Levels 1 through 6	51
3.3	Levels 7 through 12	52
4.1	Experiment Summary from Intro.	76
4.2	Instruments Used by Experiment	77
5.1	Experiment Summary from Intro.	97
5.2	Results from the first experiment.	103
5.3	Results from the second experiment.	104
5.4	Top ten dimensions (ranked by difference) from natural language processing using LIWC.	105
5.5	Number of players that stopped playing after each level.	106
5.6	Prediction accuracy of various machine learning algorithms. Higher means that the algorithm performed better.	107
5.7	The 1R attribute evaluation scores for each feature.	107
5.8	Players completing ≤ 1 levels.	120
5.9	Participants selecting same gender role models versus different gender role models.	120
5.10	Participants selecting same race role models versus different race role models.	121
5.11	Performance by role model types	122

5.12 Overall level completion statistics.	126
5.13 Female participant level completion statistics.	126
5.14 Example sentences	156
5.15 Overall level completion statistics.	161
5.16 Demographics	182
5.17 Final set of interests, scientists, and achievements	192
5.18 Regression properties β , R^2 , change in R^2 , F , and p from adding identifica- tion. Change statistics are marked (c). Significant results are bold.	206
A.1 Mazzy Versions.	218
D.1 CPSES	226
D.2 Performance—MANOVA Multivariate F-tests	227
D.3 Performance—Descriptive	228
D.4 Performance—Posthocs	229
D.5 Self-Efficacy—MANOVA Multivariate F-tests	230
D.6 Self-Efficacy—Descriptives	231
D.7 Self-Efficacy—Posthocs	232
D.8 GEQ—MANOVA Multivariate F-tests	233
D.9 GEQ—Descriptives	234
D.10 GEQ—Posthocs	238
D.11 PENS—MANOVA Multivariate F-tests	240
D.12 PENS—Descriptives	241
D.13 PENS—Posthocs	243
D.14 Choice Condition—Descriptives	244

Chapter 1

Introduction

1.1 Motivation

Educational technologies such as adaptive learning systems, educational games, and Massive Open Online Courses (MOOCs) have proliferated in recent years [564]. In 2015, the Entertainment Software Association (ESA) estimates that 155 million Americans play video games, 4/5 U.S. households own a device used to play video games, and 42% of Americans play video games regularly (3 hours or more per week) [148, 289]. Moreover, 97% of teachers use digital games created specifically for education, and 70% of teachers say they see increases in student engagement while using educational games [482]. Given the widespread and growing use of such technologies, which invariably involve virtual identities such as user profiles and avatars, it is important to better understand their impacts and to establish innovative and best practices [283]. This dissertation is encouraged by other research that has highlighted the urgency for more widespread computational literacy [67, 132, 304], and researchers like Veeragoudar Harrell who have argued for the importance of affective disposition towards the content and identification as a Science, Technology, Engineering, and Math (STEM) learner and doer [529]. For instance, studies show that representations of learners' social identities impact performance and engagement, e.g., via triggering stereotypes [503]. When learning occurs with virtual identities as intermediaries,

such as avatars in an educational game, it is unclear how the use of virtual identities may impact learners.

1.2 MazeStar Platform

To investigate this problem, we developed our own learning platform called MazeStar [291], containing within it both a Computer Science learning game called Mazzy and a flexible editor for students to build their own game levels. See Figure 1-1. Specifically, the MazeStar platform is a contribution along three axes:

- An **experimental setting** for studying the impacts of virtual identity and other phenomena, along with robust data tracking and a number of possibilities for virtual identity creation, with over 10,000 participants having taken part in controlled studies.
- A **computer science learning** framework which uses maze-solving to combine game play and game-making—extending to a wide array of computing concepts from basic programming like loops and conditionals, to human-computer interaction, design, and iterative prototyping, to more theoretical topics like search algorithms, all with heavily streamlined features including built-in image search and automatic website creation for sharing made games.
- A **focus on virtual identity** as a key component to students’ trajectories as computer science learners.

1.3 Experimental Overview

Taking an HCI approach, we have run studies using the MazeStar platform as a testbed for online experimentation. Our emphasis in these studies have been on user performance (their progress in mastering computer science concepts such as search, block structuring, and control structures) [225] and engagement (affective engagement, flow, and other issues related to user immersion) [56]. Our measures involve a combination of implicit (i.e.,

The MazeStar Platform

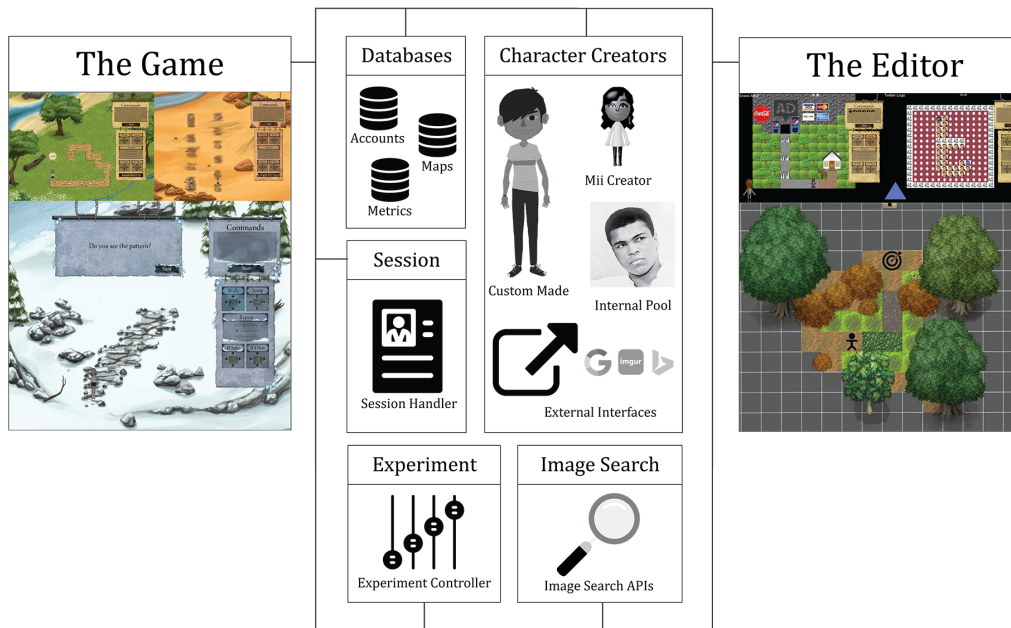


Figure 1-1: MazeStar platform components.

amount of time played, number of problems solved, etc.) and explicit (i.e., robustly validated surveys such as the Player Experience of Needs satisfaction scale [474]) measurements as well as open-ended responses. Participants in our experiments are recruited from the online crowdsourcing platform Amazon Mechanical Turk, which studies have indicated is a reliable platform for conducting experiments [360]. For reasons of experimental validity, our experiments use a between-subjects design—i.e., participants are randomly assigned to a single condition for the entire duration of the experiment. Table 1.1 has a summary of the experiments conducted along with the number of participants recruited in each experiment.

1.4 Design Principles

Here I state the design principles that have arisen as a result of this work. The following design principles will be of interest to makers of educational environments and digital contexts more generally:

1. Using avatars that **resemble users when they are doing well, and appear more minimally or abstractly otherwise**, is encouraged whenever possible. The research in this dissertation, which defines these as *successful likeness* avatars, has shown that they result in improved user performance and engagement [286]. For example, applied to a mail client or a social network like Facebook, your icon would change between a likeness of yourself or abstract depending on the positivity of your news feed, or a message you received. The essence of this principle is selectively promoting detachment and identification at key moments of the digital experience.
2. Using avatars that **resemble role models** is encouraged whenever possible. The research in this dissertation has shown that role model avatars increase both the engagement and performance of users [277, 279, 285, 290]. For example, playing as an admired and positively influential scientist, politician, business person, artist, or doctor depending on context. The criteria for an effective role model is perceived competence, similarity, and success, therefore role models should represent successful figures with demographic overlaps with users.
3. Use **embellishment** with trade-offs in mind. The research in this dissertation has shown that embellishment increases engagement, but decreases performance and self-efficacy [289]. For example, in an educational context, embellishment can be reduced to promote performance and self-efficacy, while in an entertainment context embellishment can be used more liberally.
4. Using **positive or neutral encouragement** is encouraged whenever possible. The research in this dissertation has shown that positive (e.g., “Keep it up!”, “Don’t give up!”, “You’re almost there”) and neutral (e.g., “You are doing standard work”, “You’re doing average”, “You’re doing typically”) encouragement text increases engagement [288]. For example, encouragement text can be spoken by a game character, or simply appear at the bottom of the screen periodically.

Experiment	N
Shape vs. Likeness #1	258
Shape vs. Likeness #2	250
NoAvatar vs. Likeness	182
Shape vs. Friend	208
Likeness vs. EasyLikeness	128
ScientistText vs. ShapeText	224
Shape vs. Scientist	399
Shape vs. InstantLikeness	446
Shape vs. RoleModel	357
Shape vs. Scientist vs. Athlete	1067
Phantoms vs. Non-Phantoms	523
Successful Likeness	997
Red vs. Blue	507
Feedback Positive vs. Negative vs. Neutral vs. Nothing	645
Mini-Game Loss vs. Near-Win vs. Win	366
Game Theme Basic vs. Circuit vs. RPG vs. Choice	1171
Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice	1230
Badge Type Comparison; 6 Conditions	2189

Table 1.1: Experiment Summary

- Promoting **avatar identification** is encouraged whenever possible. The research in this dissertation has shown that avatar identification promotes higher engagement, self-efficacy, time spent, and even quality of created artifacts [290]. For example, giving users the ability to customize their avatars is one simple way of increasing identification.

In the remainder of the thesis you will read about the work that led us to these principles.



(a) A sample Shape avatar.

(b) A sample Likeness avatar.

Figure 1-2: Sample avatars from the successful likeness experiment.

1.5 Findings Summary

Here, I summarize the notable findings from our experiments (for a more thorough breakdown of the data, see the [Experimental Overview](#) chapter):

Avatar-Based Outcomes:

- **Simple avatars often outperform complex avatars** [286]. This could be for a number of reasons. Seductive details [178], e.g., more complex, more embellished, etc. can be a distraction, outcome dissociation [286], e.g., non-human avatars promote less identification with failure, stereotype threat mitigation [503], e.g., simpler avatars contain fewer salient identity characteristics, and the Uncanny Valley, e.g., “almost” human avatars elicit revulsion [388].
- **Scientist role model avatars are extremely effective** [277, 279, 285]. Within a CS programming environment, all participants experience increased engagement while using scientist role model avatars, while female participants experience the most significant increases. Female participants often have significant increases in their play performance and reported engagement through using a well-known scientist as their avatar (e.g., Marie Curie), as compared to participants that used a well-known athlete as their avatar (e.g., Serena Williams), or a simple abstract shape (e.g., Triangle).
- **Successful likeness avatars can likely outperform any existing avatar types** [286].

We have discovered a new type of avatar, what we term the *successful likeness*. This is a simple abstract avatar when the user is in the trial-and-error process and a likeness of the user only when the user achieves a goal. Compared to users that used only an avatar that was always simple abstract, or always a likeness of the user, or a likeness of the user when the user was in trial-and-error and a simple abstract avatar upon achieving a goal, these successful likeness participants played significantly longer and completed significantly more levels. We propose that these results can be explained by a model in which identification facilitates vicarious outcomes, and in which detachment facilitates outcome dissociation [286].

- **Red avatars cause significant decreases in engagement and avatar affect compared to blue avatars** [287]. Research has consistently shown that red reduces mood, affect and performance in cognitive-oriented tasks [146, 190, 244, 271, 314, 329, 374, 376, 493]. For example, Lichtenfeld et. al showed that even just peripherally noticing red (e.g., hidden in a question, in the copyright notes at the end of a page, etc.) can have similar effects [329]. Prior work on first-person shooter (FPS) multiplayer games have hypothesized that blue teams are at a disadvantage because they “see red” [251]. We provide the first study to show that this effect is true in a single-player context [287]. This red-blue discrepancy was higher for male players than for female players.
- **Badges and avatar identification promote positive outcomes** [290]. We have found that badges can promote avatar identification (personal interest, role model), player experience (achievement, role model), intrinsic motivation (achievement, role model), and programming self-efficacy (role model) during both game play and game making. Independently of badges, avatar identification promotes player experience, intrinsic motivation, programming self-efficacy, and the total time spent playing and making. Avatar identification also promoted other meaningful in-editor activity, such as playtesting time, etc. and led to significantly higher overall quality of the completed game levels (as rated by 3 independent externally trained QA testers) [290].

Other Outcomes:

- **Positive and neutral encouragement text displayed at regular intervals (e.g.,**

- “Keep it up!”), significantly increases engagement as compared to no text or negative encouragement text [288].** Encouragement is different from feedback, in that it does not necessarily encode information about performance [303, 384, 444, 478]. Regularly dispensed encouragement, operationalized as text appearing at the bottom of the screen—both positive (e.g., “You’re doing good”) and neutral (e.g., “You’re doing average”) significantly increased player engagement as compared to negative (e.g., “You’re doing badly”) or none.
- **More embellished game backgrounds cause players to have significantly decreased game performance and significantly decreased programming self-efficacy but significantly increased engagement [289].** Research suggests that the addition of seductive visual details in video games hinders performance of learners [178, 455, 513]. Yet, other research results propose the opposite: that visual embellishments and well-designed ambiguity instead improve learners’ performance, engagement, and self-efficacy [488, 517, 554]. To shed light on this apparent contradiction, we implemented the following four game themes: 1) *Generic* theme with no embellishments (simple flat color background), 2) *Fantasy* game theme (forest, snow, and desert adventure backgrounds), 3) *STEM-oriented* theme (computer circuitry background), and 4) *Choice* (the user picks one of the previous three options). Generic condition participants had highest performance (levels) and had highest programming self-efficacy—followed by choice, fantasy game setting, circuitry. However, ordering of conditions for engagement was precisely opposite the trend for performance. These are trade-offs between two diametrically opposed approaches to game themes and embellishment: instrumental game skins vs. thematic and deliberately embellished game skins [289].

1.6 Generalizability

Our measures from assessing the impacts of virtual identity (performance, engagement, and identification) might also apply to other tasks. Our work has implications to gener-

alize beyond education settings to any task that may involve using a virtual identity, e.g., social networks like Facebook or LinkedIn, virtual reality simulations for educational and entertainment purposes, military virtual control enabling applications, remote teleconferencing virtually and/or robotically, conversational agents [84, 85] and other autonomous agents [347], and so on. Our work lies at the intersections of human-computer interaction, computer-supported collaborative learning, and games engineering. This thesis will follow with these core sections: 1) the Related Work in which I shall describe related research, 2) the MazeStar Platform where I describe our system and its individual constituents, 3) the Experimental Overview in which I shall describe the general experimental approaches taken, 4) the Experiments where I shall discuss our experiments in detail, 5) the Conclusion in which I shall have concluding remarks.

Chapter 2

Related Work

This work studying the impacts of avatars in computational learning environments builds on work from avatars, pedagogical agents, cognitive science and stereotyping, constructionism and computational thinking, adaptive learning, and various other strands of psychology, human-computer interaction, and artificial intelligence. This chapter summarizes relevant prior work.

2.1 AIR Project

This builds upon the Advanced Identity Representation (AIR) project [218], which constitutes approaches to analyzing and designing social categorization systems across diverse forms of virtual identity ranging from avatars to social media profiles. It is grounded in approaches to cognitive categorization and social classification from cognitive linguistics and sociology, along with HCI approaches for implementing and evaluating results. Harrell describes six components that exist in the majority of computational identity technologies (see Figure 2-1).

Identifying these components serves to both provide an appropriate level of abstraction to analyze representations across different applications, and also to identify components

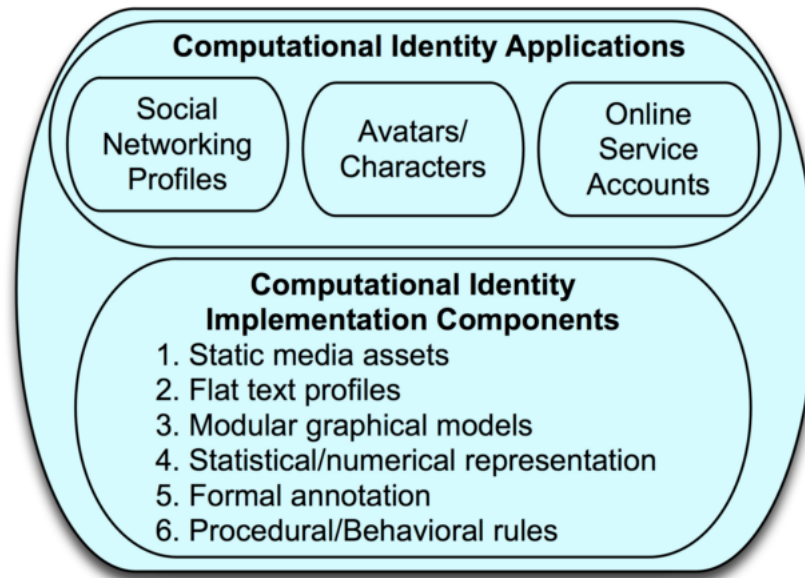


Figure 2-1: Basic components of computational identity applications.

that can be analyzed both in terms of how they appear visually and are implemented in the back-end. The AIR project identifies several important limitations found in popular computational systems, such as games (e.g., “Attributes are reduced to statistics”), social networks (e.g., “Community membership is a binary model”), and virtual worlds (e.g., “Identity representations do not express becoming or mixing”). These limitations are often ones that exist in most popular systems whether commercial, noncommercial, educational, etc. The AIR project is one lens through which we can begin to critically analyze these media.

2.2 Blended Identities

Harrell describes digital self-representations as selective projections of some aspects of a real player (e.g., preferences, control, appearance, personality, understanding of social categories, etc.) onto the actual implemented, virtual, representation [219]. As such, Harrell’s notion of a “blended identity” is an approach based on looking at structural mappings from one domain to another that is central to the understanding of virtual identities in this project

[218]. This concept builds upon James Gee’s notion of the “projective identity,” which can be described as “manifesting the ways that real player values are reconciled with values understood as being associated with avatars.” [183, 222]. This concept also relies upon metaphor theory [315] and conceptual blending theory [154]. Relating in-game behavior to real-world identities, such as demographic segments [330, 331] has demonstrated useful insight into understanding how to match interaction mechanisms in digital media systems such as games to users in order to provide the most appropriate supports. Such supports can have strong impacts on user behaviors, such as has been shown by research on the “proteus effect,” a phenomenon in which users conform to expected behaviors and attitudes associated with an avatar’s appearance [560]. For example, two of the earliest studies found that participants with taller avatars were more aggressive, and that participants with more attractive avatars were more confident. Here, our focus is on matching avatar uses with supports for computer science learning by diverse players.

2.3 Identification and Similarity

There is an abundance of work both in studying avatars and on pedagogical agents (i.e., virtual pedagogical agents, teaching agents, etc.) that guides our work. In particular, a large body of work has shown that avatars and agents that share users’ external characteristics (e.g., age, gender, race, clothing, etc.) are more influential and are linked to better learning outcomes [18, 22, 32, 202, 265, 299, 438, 464]. This is posited to be a result of similarity-attraction, the theory that people are attracted to similar others [79, 256]. Functional neuroimaging has found that perceived similarity is an important factor in a person’s ability to simulate the internal state of another person [378]. Likewise, Mobbs et. al found that when a participant watched a game show contestant with high perceived similarity, the participant experienced significant increases in both subjective and neural responses to vicarious reward [380]. Other work suggests that what is experienced by an avatar is also experienced by its user [83, 243, 386, 546, 558]. This effect is more powerful via avatars that we identify with [141, 528], identification being positively correlated to such factors as representation of

emotions and intent [212], physical resemblance [346], and avatar customization [520]. For instance, Birk, Atkins, Bowey and Mandryk, divided participants into two groups, one that customized their avatar and another that watched a video of someone customizing an avatar. Those participants that customized their avatar had increased identification. Furthermore, participants' identification with their avatars significantly predicted various measures related to engagement such as affect, immersion, and amount of time playing [47].

2.4 Stereotyping

Avatars can be pivotal in enabling our capacities to put ourselves inside other identities. However, the unfortunate consequence is avatars can also be used to reinforce stereotypes and perpetuate hegemonic views, e.g., women as victims of violence. For example, stereotype threat is the risk of confirming, as self-characteristic, a negative stereotype about one's group and originates from a study by Steele and Aronson [503]. Stereotype threat has been studied in relation to avatars [283, 445]. For example, Ratan and Sah showed that participants using a customized male avatar had higher subsequent math performance than participants using a customized female avatar [445].

2.4.1 Stereotype Threat

Blue Eyes/Brown Eyes Exercise

A good illustration of stereotype threat is Jane Elliott's well-known exercise involving eye color and brown collars. As an exercise to demonstrate racial segregation to her class, Elliott based the exercise on eye color rather than skin color. For the first day of the exercise, the blue-eyed children were designated the superior group. The brown-eyed children were asked to wear brown fabric collars such that they could easily be identified. Blue-eyed children were given extra privileges such as second helpings at lunch, five more minutes of recess, etc. Furthermore, the blue-eyed children were encouraged to play only with the

blue-eyed classmates and to ignore the brown-eyed ones. The “superior” group became arrogant, their grades were better, they completed math and reading tasks beyond their previous abilities. The “inferior” group scored worse on tests and isolated themselves. These stigmatized students barely paid attention in class. They only spoke when they were spoken to. They responded slowly and lethargically. Yet, when the exercise was reversed, such that the blue-eyed students became the stigmatized group wearing the collars, those same brown-eyed students responded with exuberance and adeptness—their status appeared to be “an actual component of their ability” [504].

Minimal Conditions for Bias

Henri Tajfel demonstrated the “minimal group effect”—that the minimal conditions for favoritism is categorization into a group, no matter how arbitrary the criteria for that categorization [511]. In one well-known experiment, 64 boys were asked to quickly judge the number of dots flashed on a screen. They were then supposedly classified as “overestimators” or “underestimators” depending on their counts—in actuality these labels were assigned completely at random. Each boy was then asked to assign small amounts of money to two other boys—when both of the boys were from the same group as the one assigning the money, they allocated the money as equally as possible. But when they were allocating between two boys, one from within their own “estimator” group and the other from the other “estimator” group, they unfailingly favored their own group [511].

No One Is Unaffected by Stereotypes

Stereotype threat has been typically framed as a phenomenon affecting currently under-represented groups, particularly in the United States where many of these studies took place, and especially women and African Americans. In reality, stereotype threat affects anyone from white males [16], Asian Americans [495], European Americans [510], men on a test of social sensitivity [306], French students [109], older Americans [232], etc. Merely the physical presence of other people of the same or different social category evokes or

suppresses stereotype threat [252]. Stereotype threat appears to occur more strongly when the individual identifies strongly with the stereotyped group. However, stereotype threat generally has but one requirement—that the person care about their performance in the stereotyped domain.

Physiological Effects

The exact physiological processes underlying stereotype threat appear to stem from stress arousal, performance monitoring, and efforts to suppress negative emotions and thoughts [485]. Stereotype threat disrupts working memory and executive function [253, 484], heart rate [110], blood pressure [53], arousal [40], self-consciousness about one's performance [36], and cause a suppression of negative emotions like anxiety [264]. When a large amount of cognitive resources are spent worrying about and ruminating on various facets related to performance pressure, individuals tend to perform worse on the task.

Threat Mitigation

Some approaches that have been found to mitigate against stereotype threat include persuading the participant that intelligence is malleable and can be improved through effort [17, 194], self-affirmations or writing about values of importance to oneself [100], giving participants a sense of social belonging within the social group [535], etc.

Role models are another avenue for reducing stereotype threat [78, 93, 139, 340, 341, 357, 372]. In one such study, participants read anywhere between 0-4 biographies of successful women. All the participants then took a difficult math test. The female participants that did not read any biographies performed worse than men. However, the more biographies that female participants read, the better they performed. Those female participants that had read four biographies performed at the same level on the math test as the men [372, 373].

There are three factors that can increase a role model's effectiveness. The first is the perception of the role model as competent [358]. The second is sharing common attributes

such as gender or race, since they are seen as an ingroup member that has overcome stereotypes [340, 357]. The third is that the role model should have achieved success [78]. One aspect studied in this thesis is role model avatars [277, 279, 285].

Criticism

Sackett, Hardison, and Cullen report that Steele and Aronson's original findings have been widely misinterpreted—many popular presses, scientific articles, and psychology textbooks cite Steele and Aronson's article as evidence that eliminating stereotype threat completely eliminates the African American–White test score gap [476]. Sackett et. al indicate that with stereotype threat removed in Steele and Aronson's study, an achievement gap of approximately one standard deviation remains. Steele and Aronson state that it is a misinterpretation of their study to cite that eliminating stereotype threat eliminates completely the test score gap [477].

Others argue that the evidence for stereotype threat is very weak, and that the level of enthusiasm is not commensurate with the evidence and thus could hamper the potential for effective interventions [508]. Some have suggested that the field of stereotype threat is inflated by publication bias [163]—many of the studies that do not find positive effects are left unpublished. Critics have argued this to be a serious concern if the results are being used to make recommendations for interventions [177]. Critics have mainly questioned stereotype threat's generalizability and the extent to which it truly explains differences in academic achievement contexts. Jussim questions whether its widespread acceptance is partly due to a narrative of egalitarianism that is “professionally risky to challenge” [272].

Overall, there is a fair amount of criticism being leveled at stereotype threat. However, there are equally large amounts of evidence for stereotype threat, e.g., replications in different domains [504]. Stereotype threat exists—however, its actual effect size is debated.

2.5 Avatar Impacts on Engagement and Performance

To the best of our knowledge, there has not been extensive work on the impacts of avatars on player engagement and performance. Linebarger et. al compared four avatar types on task performance in a virtual environment and concluded that “simpler, less computationally expensive avatar representations are quite adequate” [336]. More recently, Domínguez et. al explored the impact of avatar color on performance in a virtual scavenger hunt, although their results are so far “not conclusive” [134].

2.6 Constructionism

Constructionism is a theory of learning in which learners construct mental models for understanding the world. Cornerstones of this theory include student-based discovery learning, whereby students learn via bridges to their pre-existing knowledge and learning through production of shared artifacts [416]. *Constructivism* is a separate theory of learning from which constructionism is derived [431].

2.6.1 Constructionism’s Beginnings

Seymour Papert said of learning that it “happens especially felicitously in a context where the learner is consciously engaged in constructing a public entity, whether it’s a sand castle on the beach or a theory of the universe” [416]. Papert felt strongly that the “instructionist” approach towards education (similar to what Freire would term a “banking” concept of education [168]), which involved explicit verbal instruction, was a deficient educational approach. He said that a person can make two types of (scientific) claims about constructionism. The weak claim is that constructionism suits some learners better than traditional instructional approaches. The strong claim is that it is better for *everyone*.

In the seminal book *Mindstorms*, Papert describes “Turtle Geometry”—an environment for

programming an icon of a turtle trailing lines across a computer display—as drawing upon the child’s pre-existing pleasure and knowledge of motion. This helps provide one bridge towards more formalized notions of geometry [413]. Piaget’s constructivism demonstrated that children learn fundamental mathematical ideas through their own “false” theories first, such as preconservationist mathematics. Papert describes these “false” theories as being fundamentally necessary to one’s learning path, whereas these theories would be rejected outright in school. Rather, children’s unorthodox theories are not deficiencies or cognitive gaps, but rather ways of “flexing cognitive muscles” and working through towards more skillful orthodox understandings [413]. One of the questions Piaget asked children was “What makes the wind?” Most children gave their own theories, like “The trees. I saw them waving their arms” [415]. Despite the incorrect theory, Papert indicated that this was still demonstrative of skill in theory building—a strong correlation does indeed exist between the wind and tree branches waving [413].

Papert described early experience with “Turtle Geometry” as a good way to “get to know” more formalized subjects through some of its powerful ideas [413]. This relates to what Lave and Wenger term “legitimate peripheral participation” [318], what Crowley and Jacobs consider “islands of expertise” [111], and what Shaffer terms an “epistemic frame” [491]—all of which describe how beginners can slowly become experts, with their expertise extending far beyond the boundaries and consequences of the original activities. Constructionism places a heavy emphasis on breaking knowledge up into “mind-size” bites—similar to James Gee’s “incremental principle” [182]—making knowledge more communicable, assimilable and “constructable” [413]. Almost three decades later, Papert’s original ideas on constructionism remain relevant and have become ubiquitous in how learning theorists and educators aim to revamp traditional teaching methods.

2.6.2 Constructionism in the Present

Blikstein argues that “making” in education is becoming democratized—tasks and skills previously only available to experts have become widely learnable [55]. Most notable

are programming environments like Scratch [452] and NetLogo [551] that have brought coding to millions of students. For physical-based artifacts, Blickstein describes the rise of the “digital fabrication lab,” tracing a lineage spanning the Lego Mindstorms kit and the Cricket [353, 354] (a programmable creature) to recent developments like student-created interactive textiles [74, 75], interactive 3D worlds [105], cybernetic creatures [442, 487], etc. He argues that a school that values sports will build a gym and basketball court, a school that values music will build a music room, but that no such analagous space exists in terms of digital “making” [55].

Assessment has traditionally been a challenge in constructionist environments due to the number of variables that go into the process of constructing an artifact [54]. Philosophically, there may be valid counter-arguments as to the effectiveness of existing assessments (or even if assessment is necessary), however, researchers have argued that some alignment with assessment is necessary to prompt wider-scale policy changes [42]. Berland argues that “educational data mining” or “learning analytics” can help provide quantitative insights into constructionist environments without abandoning its richness [42]. Researchers view constructionist environments as an area of tremendous potential. Ito has characterized software into three categories: academic, entertainment, and construction, with *construction* having the most profound influence in that they allow the authoring and remixing of “media worlds” [257].

2.6.3 Modding and Constructionism

In games, we are witnessing a veritable rise of videogames and virtual environments that could be considered “constructionist” platforms. *Skyrim*, *Minecraft*, and *LittleBigPlanet 3* are all games with “modding,” taken up by gamers with enthusiasm, e.g., [437]—as of September 2016, *Skyrim* has more than 28,000 modifications on Valve’s Steam Workshop. *Counter-Strike*, *Team Fortress*, *League of Legends*, and *Dota 2* are all games that themselves are direct descendants of “mods.” Games like *Roblox* marketed for children and teenagers aged 8-18, which has—as of July 2016—15 million monthly active users [460], allow

users to texture their avatars from scratch. The next generation of gamers will grow up expecting an unparalleled control over virtual identities. *Dota 2* allows hobbyist modelers to create custom items, effects, and textures for the more than 110 heroes available [176]. Items that are accepted through Steam Workshop—a curated process involving both Valve and crowdsourced ratings [524]—nets the original modeler 25% of all revenue [396]; successful modelers receive \$3,000 - \$6,000 USD per month. While the monetary rewards are driving the number and expertise of the modelers upwards [533], modelers also cite other motivational factors: creative skills improvement, peer recognition, etc. [396]. “Modding” is not new—it has existed since the 1980s [436]—but gradually systems and processes have been put in place by developers to both lower the barrier to entry and to incentivize the act of building. Games like *StarCraft*, *Warcraft*, *Trackmania* [421]—and countless others—all shipped with official “level” editors, and could be reskinned using either official or unofficial tools. The *Sims*, *Whyville*, *Second Life* [221], all have a significant “meta-game” around making “things” for an avatar, e.g., “face-parts” in *Whyville* [275], animated textures in *Second Life* [475], clothes in *The Sims* [231], etc. This constructionist trend is also occurring in Computer Science education, e.g., [67, 75, 291]. However, the Steam Workshop is currently *the* platform for user-driven content, supporting—as of 2016—almost 500 titles [525]—with adoption of the platform even being predictive of higher user ratings [305].

2.6.4 Instantaneous Selves

In the near future, the barrier to entry for “modding” will be non-existent. The technology is here—it is now possible to create a fully rigged, personalized 3D facial avatar from hand-held video—i.e., from a cell-phone camera—that can be animated in realtime [248]. Scanning our bodies, and any material objects, into a virtual world will require only standard consumer devices. Developing high-quality 3D hair models from a single 2D image can now be done [87]—we can imagine the vast online world of 2D images that can be quickly transformed to their 3D counterparts. Example-based stylization [161], “wear-and-tearing” arbitrary textures [37], complex real time hair simulations [86], are just some of the examples

that will further lower the barrier to entry to “modding”. As suggested in [284], automatic avatar generation in a specific art-style—with specific alterations—is possible right now.

2.7 Computational Thinking

Computational thinking is most widely understood through Cuny, Snyder, and Wing’s definition [555]:

Computational Thinking is the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent.

Historically, computational thinking was a term first used by Seymour Papert in 1980 [412, 414], and in the ensuing decades has taken on different aliases albeit with philosophically similar definitions—computational literacy, which focused more on computing as a medium for exploration [131], and procedural literacy, which focused more on computational thinking in the context of new media art and design [58, 361, 492]. Here, I review the most widely understood definitions of computational thinking, and mention some of the ongoing criticism.

2.7.1 What Is Computational Thinking?

While the precise definition for computational thinking has existed in various forms [9, 201, 466, 555], the general consensus is that it represents thinking like a computer scientist when confronted with a problem [201]. In 2017, the AP Computer Science Principles course was launched to over 2,700 high schools and over 45,000 students [103], the largest AP course launch to date. Developed by the College Board and the National Science Foundation, CS Principles has the following seven organizing principles:

1. Computing is a creative human activity that engenders innovation and promotes exploration

2. Abstraction reduces information and detail to focus on concepts relevant to understanding and solving problems
3. Data and information facilitate the creation of knowledge
4. Algorithms are tools for developing and expressing solutions to computational problems
5. Programming is a creative process that produces computational artifacts
6. Digital devices, systems, and the networks that interconnect them enable and foster computational approaches to solving problems
7. Computing enables innovation in other fields including science, social science, humanities, arts, medicine, engineering business

According to Grover and Pea [201], the following elements are generally accepted as being foundational to computational thinking:

- Abstractions and pattern generalization (including models and simulations)
- Systemic processing of information
- Symbol systems and representations
- Algorithmic notions of flow of control
- Structured problem decomposition (modularizing)
- Iterative, recursive, and parallel thinking
- Conditional logic
- Efficiency and performance constraints
- Debugging and systematic error detection

More specifically, in a context of design-based activities in Scratch, Brennan and Resnick define their own computational thinking framework [67]:

- Computational Concepts
 - Sequences
 - Loops
 - Events
 - Parallelism

- Conditionals
- Operators
- Data
- Computational Practices
 - Being incremental and iterative
 - Testing and debugging
 - Reusing and remixing
 - Abstracting and modularizing
- Computational Perspectives
 - Expressing
 - Connecting
 - Questioning

While various frameworks exist for defining computational thinking, there is universal agreement that computational thinking is broadly important in virtually all subject areas: as varying as biology, astronomy, archaeology, chemistry, economics, journalism, law, medicine and healthcare, meteorology, neuroscience, sports, etc. [555]. The broad applicability of computational thinking is perhaps most aptly described by Wing [555]:

Consider these everyday examples: When your daughter goes to school in the morning, she puts in her backpack the things she needs for the day; that's prefetching and caching. When your son loses his mittens, you suggest he retrace his steps; that's backtracking. At what point do you stop renting skis and buy yourself a pair?; that's online algorithms. Which line do you stand in at the supermarket?; that's performance modeling for multi-server systems. Why does your telephone still work during a power outage?; that's independence of failure and redundancy in design. How do Completely Automated Public Turing Test(s) to Tell Computers and Humans Apart, or CAPTCHAs, authenticate humans?; that's exploiting the difficulty of solving hard AI problems to foil computing agents.

Wing says of computational thinking that “Ubiquitous computing is to today as computational thinking is to tomorrow. Ubiquitous computing was yesterday’s dream that became today’s reality; computational thinking is tomorrow’s reality” [555].

2.7.2 Initiatives

There is widespread support for making computational thinking more prevalent from industry (e.g., Google, Microsoft, etc.), government (e.g., NSF, U.S. Congress, etc.), and international efforts more broadly (e.g., UK’s British Royal Society, universities and other organizations worldwide, etc.). There are numerous K–12 efforts for strengthening computational thinking such as the Georgia Computes! alliance, the AP CS Principles course, and the Exploring Computer Science curriculum. Moreover, there are countless environments and tools that support computational thinking, e.g., Scratch, Alice, MIT App Inventor, NetLogo, Game Maker, Kodu, the Arduino, etc. But despite the widespread agreement that computational thinking will be crucial to the world’s ecology, there are numerous debates and criticisms [28, 201].

2.7.3 Criticism

Computational thinking has been criticized for having multiple interpretations, for being vague, and for lack of clarity on what CS is as a discipline [201]. For example, whether computational thinking should be incorporated into education as a general topic, a discipline-specific topic, or a multidisciplinary topic [107]. As a result, some researchers have developed more discipline-constrained models of computational thinking [541]. There is even the question of whether computational thinking is sufficiently different from other types of thinking that humans develop as a matter of course. But advocates of computational thinking posit that computational thinking uniquely has as a central focus “information processes” [120, 201, 555]. Likely the most long-standing issue—and not only of computational thinking but educational environments more broadly—is that of transfer. How can one evaluate

computational thinking, and its transferability? For example, “Now that the student can program Space Invaders, can the student program a science simulation?” [201, 308]. These ongoing contentions will continue to shape the future of computational thinking research.

2.8 Adaptive Learning

Adaptive learning/intelligent tutors have been part of the AI movement since the 1970s. Adaptive learning has sought to improve and make more effective the learning experience for users. Generally, models of adaptive learning systems involve:

- Expert Model (the actual information to teach)
- Student Model (information about the student)
- Instructional Model (how the information is taught)
- Instructional Environment (the user interface)

However, work has heavily focused on the manipulation and understanding of content. Adaptive highlighting of textbook content [371], integration of content into social networks [494], focusing on why students answer certain problems incorrectly (e.g., [537]), are a few examples. Or, when there has been a focus on the student, it has been in the way of learning styles [90, 158] or other knowledge/concept-based models. Surprisingly, very few systems have endeavored to model identity (social/virtual)—this is despite their clear importance in the literature [503, 560]. Work in education has only been tangential to the topic of identity, such as virtual agents that share our gender [32], detecting students’ emotional state [197], and questionnaires seeking to model students learning through self-efficacy (“even if the work is hard, I can learn it”), etc. [377].

2.9 Other Systems/Games That Teach Computer Science

Other games and systems have been used to teach programming and/or CS principles. Non-exhaustively, these include the Logo programming language and associated turtle graphics [343], the Scratch environment [452], Alice [105] and Storytelling Alice [294], NetLogo [550], MIT App Inventor [557], Gidget [319], LightBot [2], CodeCombat [3], BOTS [233], RoboBuilder [540], Greenfoot [309], AgentSheets and AgentCubes [449], Code.org exercises [99], the Arduino [75], Kodu Game Lab [509], Game Maker [98, 407], Gogo Boards [498], the STELLA programming language [308], and others [226].

Chapter 3

The MazeStar Platform

In this section, I describe the MazeStar platform in more depth. See Figure 3-1 for an overview. I begin by describing the game (Mazzy) and its components, then the editor and its components, and finally the shared components between the two.

3.1 The Game

The experiments take place in a STEM learning game called Mazzy [278]. Mazzy is a game in which players solve levels by creating short computer programs. In total, there are 12 levels in this version of Mazzy. Levels 1-5 require only basic commands. Levels 6-9 require using loops. Levels 10-12 require using all preceding commands in addition to conditionals. See Figures 3-2 and 3-3.

3.1.1 Command Types

Table 3.1 describes the types of commands that players use in Mazzy. As seen in Figure 3-3, the player can select from different modes, each of which correspond to one command type. Within their respective modes, all commands are input using the keyboard's W (up), A (left),

The MazeStar Platform

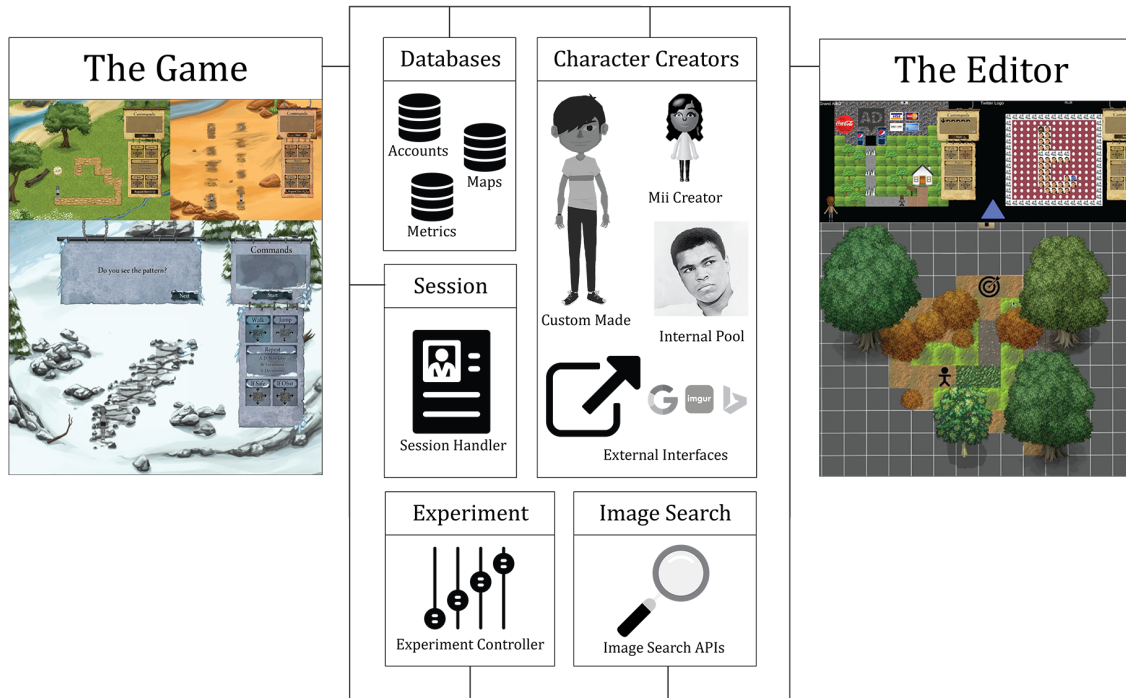


Figure 3-1: MazeStar platform components.

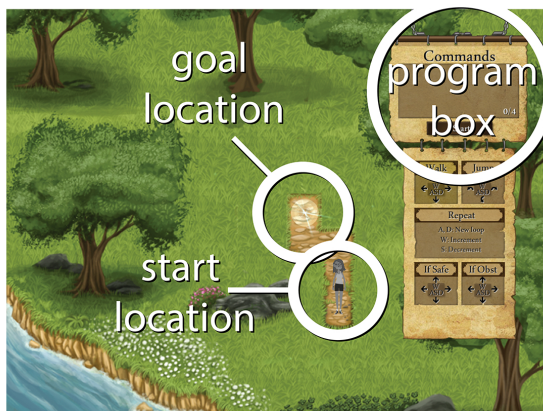


Figure 3-2: Level 1 in Mazzy introduces the basic game mechanics

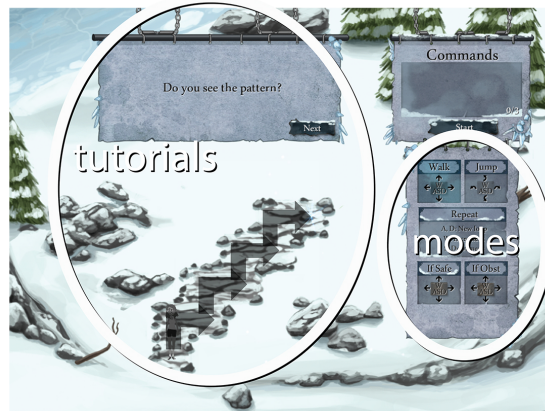


Figure 3-3: Level 6 introduces looping

S (down), and D (right) keys. This maps directly to the directionality specified in each of the command types, with the exception of the repeat command. The repeat command is created with the A, or D keys. Number of iterations is increased with W, and decreased with S, with the input caret in between the beginning and ending symbols. If there are multiple nested loops while incrementing and/or decrementing, the innermost loop with respect to the input caret is modified.

All commands can be deleted in batch by highlighting and using the keyboard's delete key—that is, commands are treated in a similar way to text within a text editor. In the case of the repeat, if safe, and if obst, commands, should only an opening bracket or closing bracket be caught in a user delete, the corresponding other bracket will also be removed. Because both the instantiation of these command types and the deletion of these command types always involves the opening and closing bracket, it is not possible at any time to have a beginning or ending bracket without the other.

3.1.2 Command Limits and Compilation Errors

By design, compilation bugs, run-time errors, infinitely executing programs, etc. are not possible in Mazzy. A program will run with a blank program, a single empty conditional, a loop with 0 iterations, etc. In these cases, the player character remains stationary. However, every level has a command limit. This is effectively a maximum number of commands to constrain the possible solution set for a given level. While many of the later levels involving loops could be solved just by using enough walk and jump commands—command limits increase game difficulty and enforce usage of the more complex command types. The player is unable to run their program if they pass a level's command limit. In Mazzy, all commands count as 1 command, except the jump command which counts as 2.




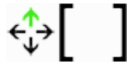
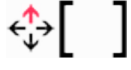
Command	Symbol	Effect
Move		Moves the character one space in the direction specified.
Jump		Moves the character two spaces in the direction specified. The first space can be an obstacle.
Repeat		Loops commands in between starting and end symbols a specified # of times. Can be nested.
If Safe		Conditional check if the space in the direction specified is safe. Successful check means commands in brackets will run. Can be nested.
If Obst		Conditional check if the space in the direction specified is an obstacle. Successful check means commands in brackets will run. Can be nested.

Table 3.1: Mazzy Commands

3.1.3 Level Overview

See Tables 3.2 and 3.3 for screenshots of each Mazzy level as well as sample possible solutions. For a video of completing these levels, see <http://youtu.be/n2rR1CtVal8>.

3.1.4 Correct Solutions

In Mazzy, a program is correct if the character (or all characters, in the later levels) reach a goal space within the command limit. Characters stop executing the program the moment they land on a goal space. With multiple characters in a single level, each character independently executes its own version of the program in parallel.

Generally, there are many correct solutions. A jump command can always be used in place of two consecutive walk commands in the same direction, experienced players can use loops/conditionals before they are introduced, conditionals can be flipped, loops/programs can be longer than necessary so long as the character or all characters reach the goal, many tricks exist for optimizing and shortening programs by exploiting the command

Level	Screenshot	A Sample Solution
1		
2		
3		
4		
5		
6		

Table 3.2: Levels 1 through 6






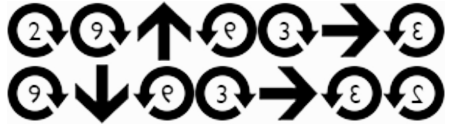



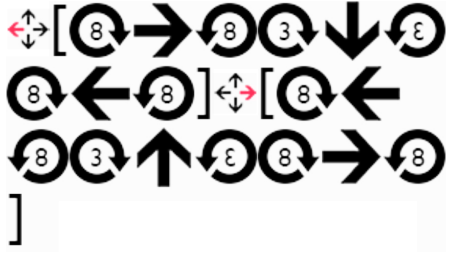


Level	Screenshot	A Sample Solution
7		
8		
9		
10		
11		
12		

Table 3.3: Levels 7 through 12

types/mechanics, etc.

3.1.5 Computing Concepts Covered

Computing concepts directly covered in Mazzy itself include sequencing commands, loops, conditionals, block structuring, and parallelism/concurrency. Players inevitably are required to repeatedly test and debug their programs, which is in line with other computational practices such as being incremental and iterative.

3.1.6 Previous Versions

Mazzy's development was an iterative process. Feedback was solicited on prototypes, as well as throughout all experiments. Here, I detail the first version of Mazzy that was used for early experiments. This version was significantly different, and is described here for completeness.

First Version

The first version of Mazzy used in experiments was prototyped over the course of 4 user evaluations (with 16, 24, 41, and 27 participants respectively). These user evaluations were a mix of both in-person participants at MIT, and online participants via Amazon Mechanical Turk and Reddit. Early feedback from users centered around desiring more instructions on how to play the game. This was in large part because early prototypes had no instructions, only just-in-time (JIT) hints that would appear depending on the user's progress. These JIT hints would periodically fade in and out to bring the user's attention to perform an action. See Figures 3-4 and 3-5. Other types of feedback that users provided related to desiring higher graphics fidelity, a larger game window that could be maximized, more challenging levels, etc.



Figure 3-4: JIT hint to click the panel



Figure 3-5: JIT hint to use arrow keys

Final Build of First Version

For a video of the final build for this first version, see <http://youtu.be/j0TI4MH2rsY>. The final build had 3 playable levels. See Figure 3-6. The first level involved simply programming the character to exit the maze. The second level involved programming three characters in parallel to each individually find the appropriate exit. The third level had fixed programs for all three characters; the user had to instead program the level to ensure that only the correct character made it to the exit. Users could still read the programs of each character in the final level, but could not change them. Each level had up to three bonus items that could be collected by the characters with more complex/difficult player solutions.

In this first Mazzy version, each level had an animated repeating tutorial. These tutorials simulated mouse clicks through a gray circle denoting a mouse click, and keyboard presses through a small keypad would appear and highlight pressed keys. Tutorials solved a conceptually similar, but simplified, level. See Figure 3-7 for still-frames. An icon could also be hovered over for text explaining the current objective and other tips.

Each level generally started with “stub” code, a program that executed but only moved the player character a few spaces in the correct direction. Upon the player executing a program, each command (arrow) would be highlighted in yellow as it was being executed. For multiple characters running concurrently, players could switch between each of the characters’ programs during execution and see the current command being executed for each individual character. This command highlighting was in effect a debugging support.



Figure 3-6: Levels in first version of Mazzy.

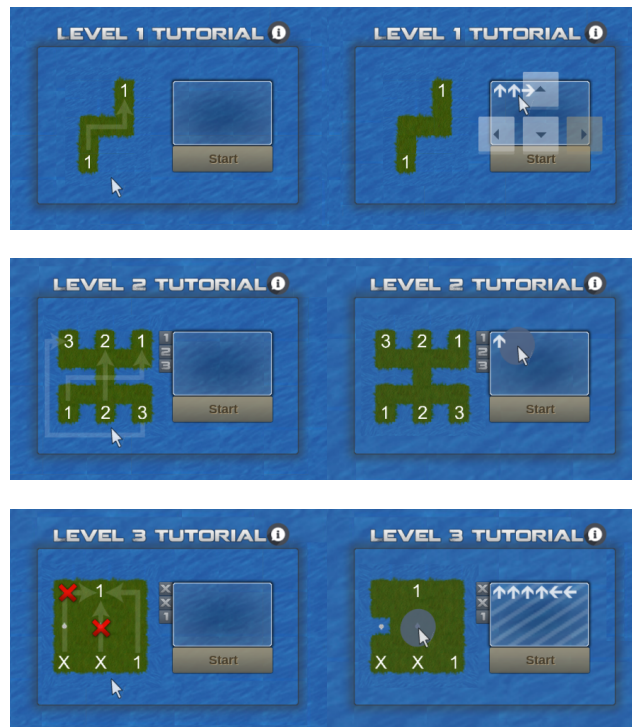


Figure 3-7: Animated tutorials.

Lastly, this first version of Mazzy could scale to any browser size to accommodate both smaller and larger/maximized displays.

Version Listing

For a full listing of Mazzy versions, the differences between them, as well as links to playable builds, see Table A.1.

3.2 Experiment Controller

The experiment controller is a sub-module of MazeStar that performs all experiment-related operations. This includes condition randomization, enabling/disabling dozens of functionalities such as in-game surveys (e.g., asking the user to rate a specific level's difficulty), controlling the avatar type of the user, controlling whether early quitting of the

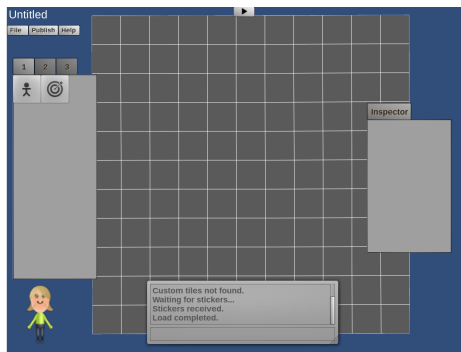


Figure 3-8: Blank 11x11 map.

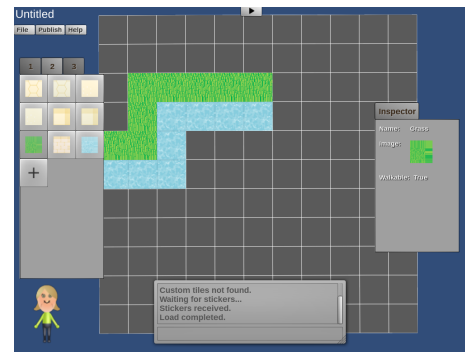


Figure 3-9: Grass and water tiles.

game is enabled, tutorials, difficulty setting, per-level experimental manipulations (such as changing the avatar when a specific condition is met in the game level), loading avatars of other players who have recently played the game and running the code they had used (as a hint-support), reskinning level backgrounds, and a host of other functionalities related to experimental set-up.

3.3 The Editor

At a high-level, the editor allows players to create their own Mazzy game levels, and then share those levels through links and automatically generated webpages. Each map consists of a grid of tiles, each of which can be textured separately and modified logically to be a safe or unsafe tile for the player to step on. The maps can be any size (from 1x1 to any size that is able to be handled in browser memory). See Figure 3-8.

3.3.1 Editor Basics

Within the editor, players move the view of the current working map using W, A, S, and D on the keyboard. They can save maps and open previously saved maps. On the left hand side is a panel that allows players to add different elements to the maps. In the first tab of this panel, players can set the start and goal location of the player (the start being where

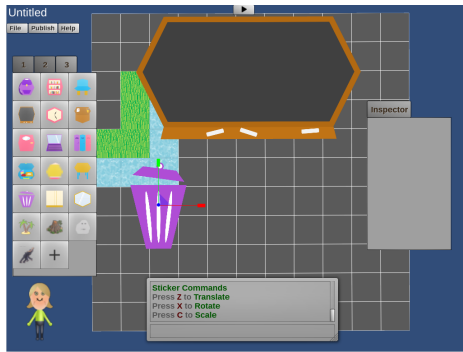


Figure 3-10: Stickers.



Figure 3-11: Searching for “cat”.

the player will initially spawn, the goal location being where they intend the player to try to reach, though the latter is not necessary for playing the map). The second tab contains textures for the tiles themselves, which can be placed on each of the grid squares. See Figure 3-9.

3.3.2 Stickers

The third tab on the left hand side panel contains textures for stickers, which are aesthetic images that appear ovetop of the grid and do not affect the game logically. These stickers can be translated, rotated, and rescaled using the Z, X, and C keys respectively to switch modes. In Figure 3-10, two stickers have been placed, with the trash bin sticker currently selected and currently in rescale mode.

3.3.3 Custom Images

Players can not only use the tiles and stickers that are pre-loaded with the editor, but also (using the plus icon shown in Figure 3-9 and Figure 3-10 at the bottom of the left-hand panel) can search for images and import them directly. See Figure 3-11 and Figure 3-12.

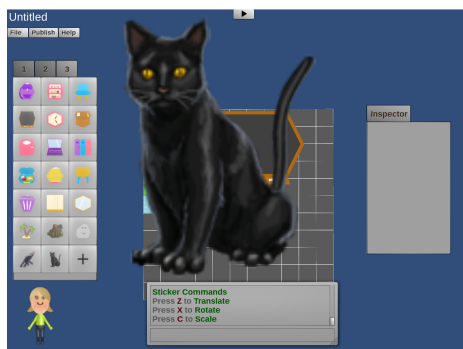


Figure 3-12: Placed search item.



Figure 3-13: Playing the “cat” map.

3.3.4 Testing a Map

Maps are periodically saved automatically to prevent data-loss in the event that the user should accidentally quit the browser without saving or in the event of a CPU crash. To test their maps, players can click on the play icon at the top-center of the screen. This simulates playing the map that they have created. See [Figure 3-13](#).

3.3.5 Sharing a Map

When satisfied with their map, players can then share their map either using: a) an automatically generated tinyURL link ([Figure 3-14](#)), or b) an automatically generated website (see [Figure 3-15](#), [Figure 3-16](#), and [Figure 3-17](#)). In the latter case, the website is a permanent record of their map and does not change (unless the user re-generated the website in which case the old one is overwritten). The automatic website generation involves the Unity program communicating with Javascript, which in turn communicates with the server using PHP. In both cases, using a link or webpage, visiting players can play the created map directly (for a visiting player, they jump straight into playing the map without needing any account credentials, going through the editor, etc.—similar to sharing a file on Google Drive or Dropbox publicly).

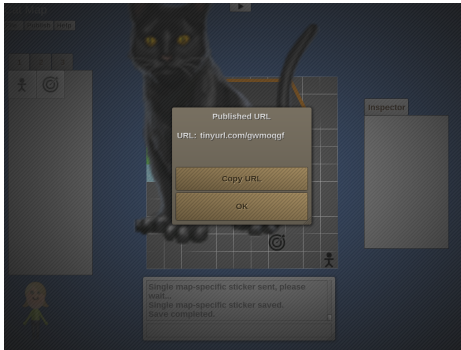


Figure 3-14: Sharing a direct link.

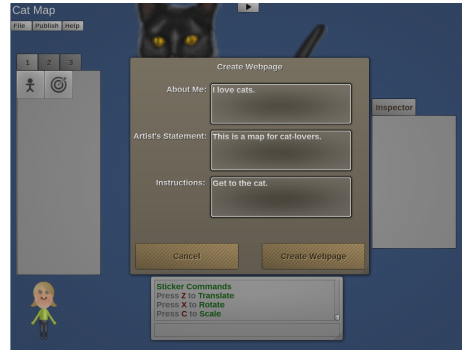


Figure 3-15: Sharing a website.



Figure 3-16: Generated webpage.

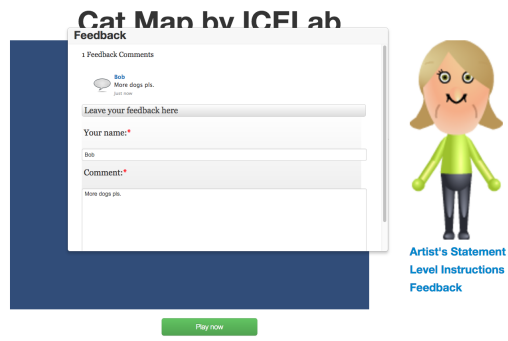


Figure 3-17: Map feedback form.

3.3.6 Example Player-Created Maps

In this section, I share 16 Amazon Mechanical Turk player-created maps. Average creation time for these 16 maps was 21.9 minutes (SD = 18.4). Players played Mazzy (for as little or as long as they liked), then were given a brief tutorial (M = 4.8 minutes, SD = 2.4) on how to use the editor. The tutorial introduced basic functionalities of the editor: panning/zooming, play-testing, searching for tiles/stickers, sticker manipulation using scaling/rotation/translation, and creating blank maps. In their version of the editor, no default images were provided for tiles/stickers (all images as part of their maps are searched for by players themselves through the editor’s image searching functionality). These maps were selected on the basis that they appeared be effective and/or creative. See Figures 3-18, 3-19, 3-20, 3-21, 3-22, 3-23, 3-24, 3-25, 3-26, 3-27, 3-28, 3-29, 3-30, 3-31, 3-32, and 3-33. The player-given map name, gender, and age are in each caption. The map shown in Figure 3-34 was instead made during a week-long workshop at a Boston high school. The student’s self-selected theme



Figure 3-18: “HomeRoad” features a variety of assets. M/29.

was a critique of social media.

3.3.7 Image Search

This component of MazeStar handles searching and importing user-selected images. Although images and avatars can also be uploaded directly from a user’s machine, this component allows users to search for images on the internet and have them displayed and imported directly into MazeStar. This is implemented using Javascript and PHP. In order to bypass crossdomain security issues related to retrieving images, a PHP script simulates an actual user to connect to and retrieve each image individually. Image type information, etc. is automatically extracted, at which point the image is converted to a byte array that is transferred back to Unity C# via Javascript. While this process is constant regardless of the external source domains of the images, it currently interfaces with Microsoft Azure (i.e., Bing).

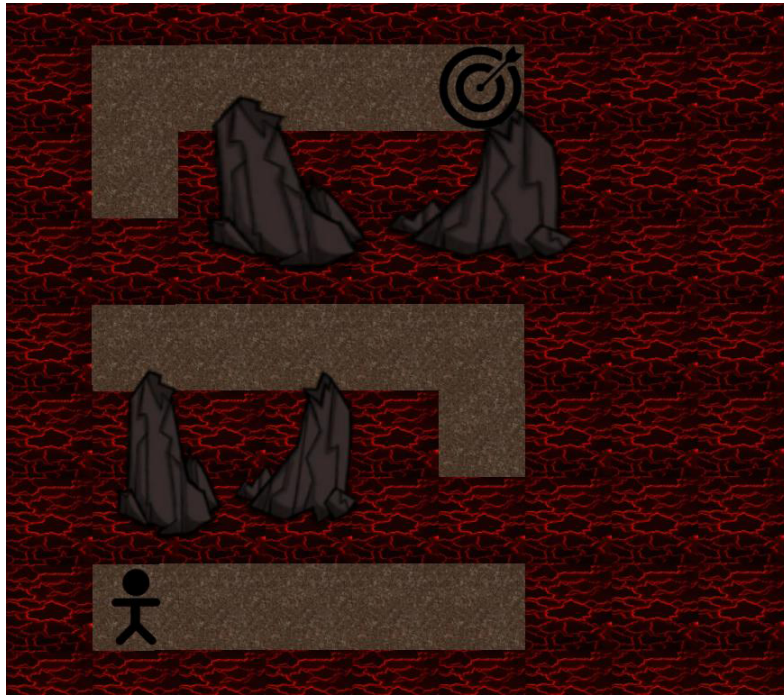


Figure 3-19: “Lava Jump” has a basic, but effective design. M/20.



Figure 3-20: “Picnic Time” creatively uses stickers to make the path “fuzzier.” F/24.

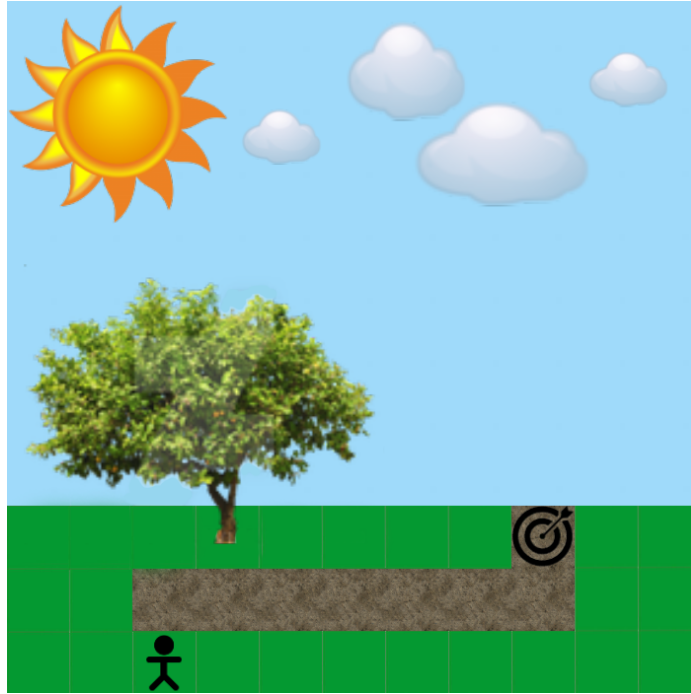


Figure 3-21: “Summer Days” creatively uses tiles/stickers to create a horizon and sky. F/36.



Figure 3-22: “The ground is Lava” is a simple and visually cohesive map. M/31.



Figure 3-23: “Starcraft theme” is . . . a starcraft theme. M/29.

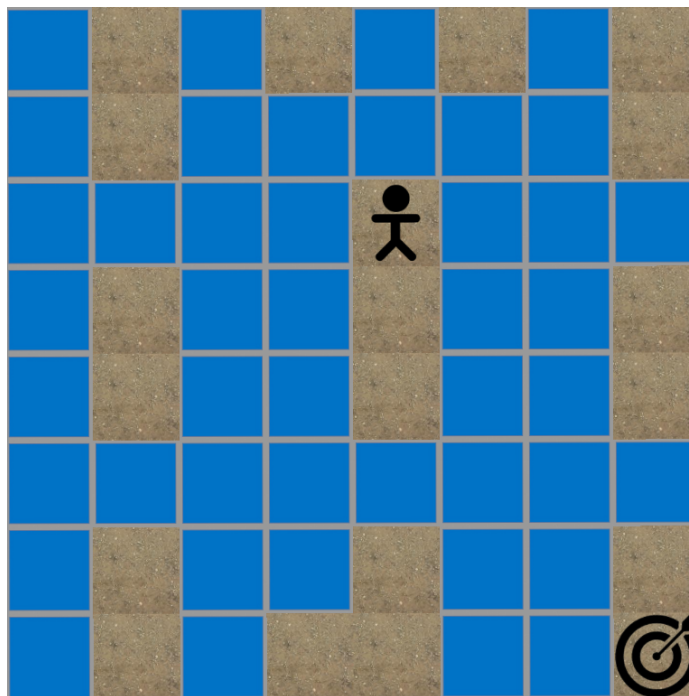


Figure 3-24: “Map1” requires going the long way. M/27.



Figure 3-25: “Island volcano” is a large 30x30 map. M/32.

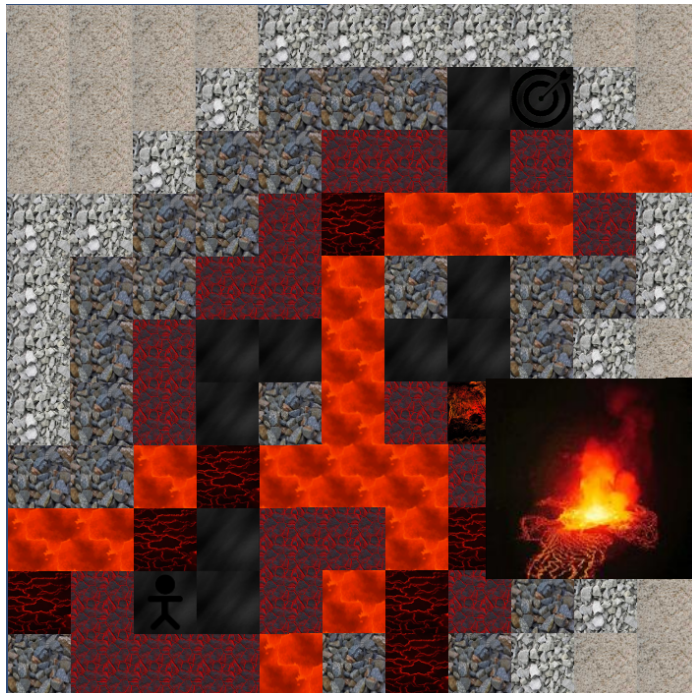


Figure 3-26: “LavaZone” uses a variety of different-colored tiles to good (gradient) effect. M/25.



Figure 3-27: "4Corners" features different habitats. M/40.

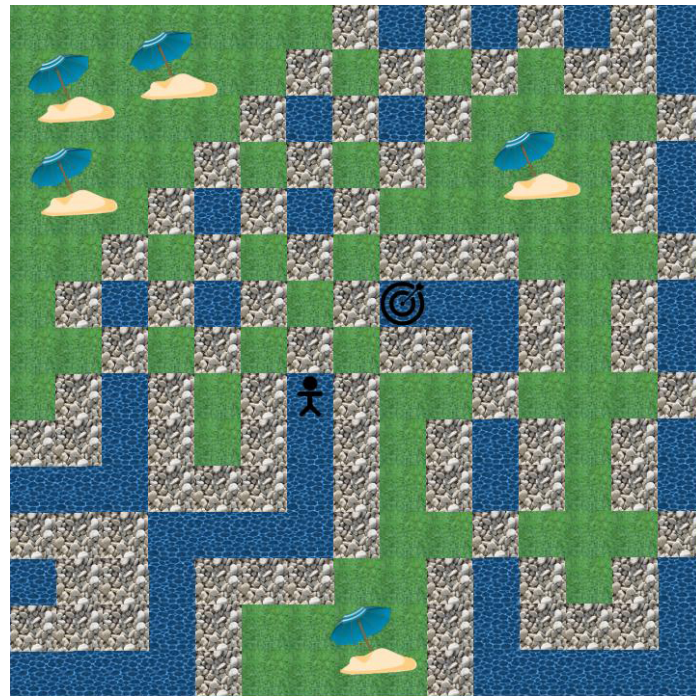


Figure 3-28: "Crossing the Stream" requires a large number of staggered jumps. F/24.



Figure 3-31: “jennymap” is also colorful, and has a “retrographics look.” F/29.



Figure 3-32: “Garden” is a long map like many scrolling games. M/21.

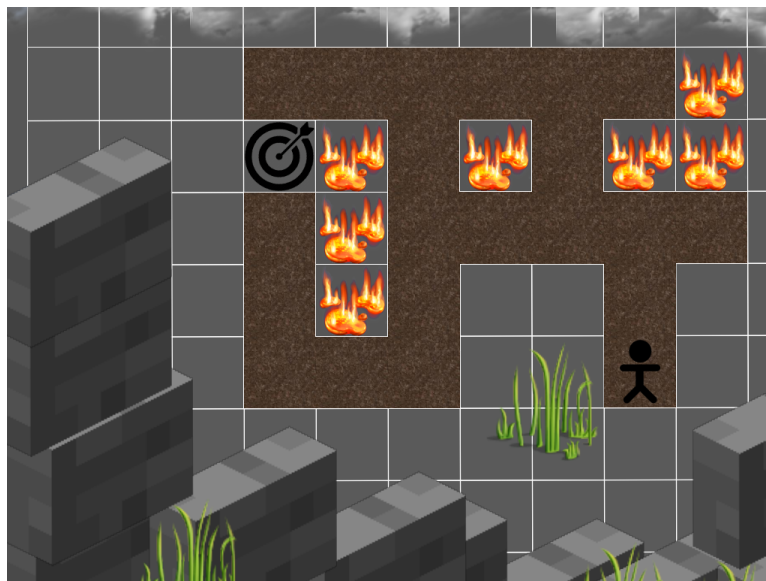


Figure 3-33: “The Ground is Lava” features an unorthodox use of perspective. F/35.

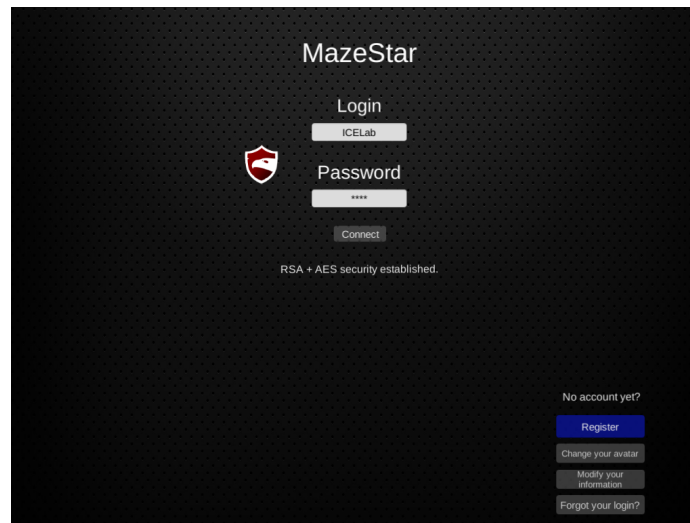


Figure 3-35: MazeStar login screen.

(associated map, x and y coordinates, etc.), actual instances of stickers (associated map, xyz location, rotation, and scale, etc.) and other metadata (start location, goal location, etc.).

Metrics Database

This database was previously used for an experiment that involved showing players “shadows” of previous players in each of the levels (to see if this could act as a type of programming help support or possibly a competition mechanic). This database contains player IDs, player avatar data (such that exact replicate avatars could be reproduced in other people’s games, including color, gender, shape, etc.), player code for each game level, timestamps, etc.

3.3.9 Session Handler

This is an internal component that tracks all pertinent data relevant to usage of the MazeStar platform. Variables such as amount of time creating or choosing an avatar, link to avatar (if the avatar was an external image), the experiment condition, experiment-specific data values (e.g., avatar name, pre-test ratings, etc.), time played, survey ratings after each level (if enabled), number of attempts, number of hints, and time taken for each level, and an

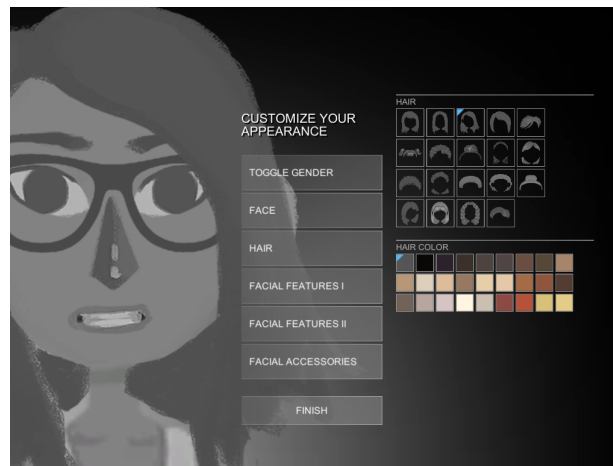


Figure 3-36: Built-in avatar creator.

array of other data tracking that can be enabled. The handler packages all the data into a single encoded string for easily exporting to, e.g., Mechanical Turk, Qualtrics, etc.

3.3.10 Avatar Creators

There are a variety of different ways that the platform's avatar creation process can take place depending on the current needs. These different ways are described here.

Custom Made

This component is a built-in avatar creator¹ that we created that can be used to create the avatar appearing both in the game and in the editor while making game levels. A variety of different facial features can be selected and colored. See Figure 3-36, Figure 3-37, and Figure 3-38.

¹These assets were used to develop an automatic portrait creator based off a user's photo that could be used for studies where an instant customized avatar was desired requiring no user intervention, though is not in live use now (<http://youtu.be/MH2Ww-r46Xc>; [284]).



Figure 3-37: Example avatar faces.



Figure 3-38: Two full-body avatar examples.

Mii Creator / AIRvatar

The Mii Creator has been used in numerous studies related to MazeStar. Moreover, recently MazeStar has been integrated with AIRvatar, a system for studying virtual identities and user behavior while creating and customizing virtual identities [332, 333]. AIRvatar/Mii Creator are currently coupled with MazeStar to make avatar creation seamless. Specifically, AIRvatar caches generated avatars into browser local storage, and MazeStar then retrieves those cached avatars, uploads them, and saves them to a user's account. See Figure 3-39.

Internal Pool

There are numerous pools of avatars already available inside MazeStar, generally used for experimental purposes (e.g., a pool of well-known scientists, athletes, abstract shapes, etc.).

External Interfaces

While avatars can be uploaded from disk (local hard drive), there exist a variety of other methods that MazeStar can use to scrape user-chosen avatar images (from imgur.com, Bing, Google, etc.).

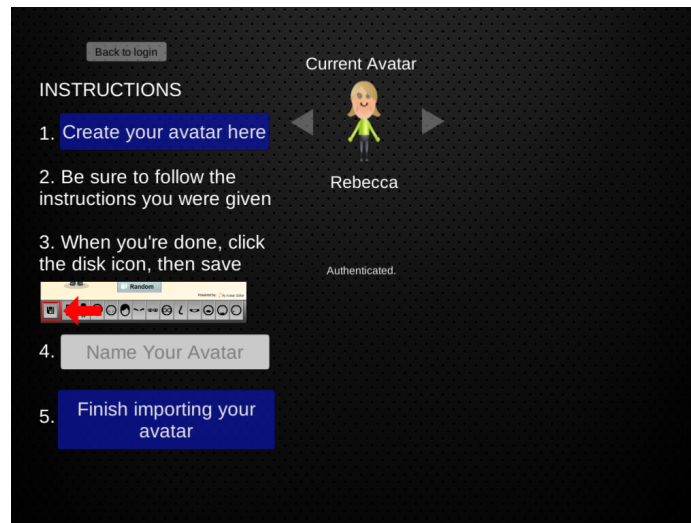


Figure 3-39: Creating a Mii.

3.3.11 Computing Concepts Covered

Computing concepts/practices encouraged through the *MazeStar* editor include quick prototyping, playtest and evaluation, and re-design stages. This is aligned with design-based models from human-computer interaction, which encourage design-create-evaluate iterative cycles (with each cycle becoming increasingly complex and higher investment). Being incremental and iterative, i.e., starting small and simple, is widely understood to be a good approach towards product-building in virtually all domains. Technological supports for this include seamless switching between playtesting/creating, sharing web links to one's game, and sharing webpages with a feedback form. Feedback on one's map might include visual appeal, fun, difficulty, clarity, etc. These are at once artistic, design, and engineering problems. Thinking through how other players might approach one's level, or about how to approach other players' levels and how they could be subsequently improved, in addition to being an integral part of iterative prototyping, also invokes concrete programming concepts. With additional facilitation, this can lead to thinking about maze-solving more abstractly/generally, such as search algorithms like breadth-first search and depth-first search.

Chapter 4

Experimental Overview

This chapter is meant to provide an overview of our experiments and work to date. If you would like a high-level overview of the work that was carried out, this is likely a good chapter to look at. If you are interested in only a specific experiment or a specific result, the next chapter looks at each experiment separately and therefore may be more suitable. However, this chapter provides additional value as it can help facilitate a higher-level understanding of our research values and aims.

I re-iterate the main findings from our studies (also previously stated in the introduction) at the end of this chapter. These findings represented what appeared to be the most consistent and powerful findings (i.e., multiple different studies confirmed it and/or it appeared to be a robust and convincing result). However, for the sake of conciseness, this subset necessarily leaves out other potentially interesting results from this high level overview, the full list of which can be found under each individual experiment in the next chapter.

4.1 Methods Overview

Table 4.1 provides a listing of the experiments that will be discussed in this thesis, along with sample sizes and designs. Here, I elaborate on our experimental methods used throughout

Experiment	N
Shape vs. Likeness #1	258
Shape vs. Likeness #2	250
NoAvatar vs. Likeness	182
Shape vs. Friend	208
Likeness vs. EasyLikeness	128
ScientistText vs. ShapeText	224
Shape vs. Scientist	399
Shape vs. InstantLikeness	446
Shape vs. RoleModel	357
Shape vs. Scientist vs. Athlete	1067
Phantoms vs. Non-Phantoms	523
Successful Likeness	997
Red vs. Blue	507
Feedback Positive vs. Negative vs. Neutral vs. Nothing	645
Mini-Game Loss vs. Near-Win vs. Win	366
Game Theme Basic vs. Circuit vs. RPG vs. Choice	1171
Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice	1230
Badge Type Comparison; 6 Conditions	2189

Table 4.1: Experiment Summary from Intro.

those experiments.

4.1.1 Measures

In this section, I describe the measurement instruments used in our studies. This is non-exhaustive, and includes ones that were recurring and/or important. Roughly, each of these are connected to the following constructs of interest:

- Player Experience of Needs Satisfaction (PENS–*engagement*)
- Game Experience Questionnaire (GEQ–*engagement*)

Experiment	Measures
Shape vs. Likeness #1	PERF
Shape vs. Likeness #2	PERF
NoAvatar vs. Likeness	PERF
Shape vs. Friend	PERF
Likeness vs. EasyLikeness	PERF
ScientistText vs. ShapeText	PERF
Shape vs. Scientist	PERF
Shape vs. InstantLikeness	PERF
Shape vs. RoleModel	PERF
Shape vs. Scientist vs. Athlete	PERF, GEQ
Phantoms vs. Non-Phantoms	PERF, GEQ
Successful Likeness	PERF, GEQ
Red vs. Blue	PERF, GEQ
Feedback Positive vs. Negative vs. Neutral vs. Nothing	PERF, GEQ
Mini-Game Loss vs. Near-Win vs. Win	PERF, GEQ, PENS, CPSES
Game Theme Basic vs. Circuit vs. RPG vs. Choice	PERF, GEQ, PENS, CPSES
Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice	PERF, GEQ, PENS, CPSES
Badge Type Comparison; 6 Conditions	PERF, PENS, CPSES, IMI, PIS

Table 4.2: Instruments Used by Experiment

- Computer Programming Self-Efficacy Scale (CPSES—*computational identity*)
- Intrinsic Motivation Inventory (IMI—*engagement*)
- Player Identification Scale (PIS—*virtual identity*)
- Performance (PERF—*performance*)

Note that each questionnaire/scale is almost inevitably more broadly applicable than the construct for which we use it for, although each of those represent our main interest in the particular instrument. Sometimes, multiple instruments from the same construct of interest are used in the same study as additional coverage/reliability. For example, engagement is fundamentally important to learning environments [56], and self-efficacy is a strong predictor of women’s career choices, especially in regards to STEM fields [45, 410, 566]. See Table 4.2 for a summary of measures used in each experiment¹. Experiments listing only performance (PERF) used other methods not listed above, e.g., in-game polls/questionnaires, etc.

¹Experiment conditions will be described later in this chapter and in the next chapter.

Player Experience of Needs Satisfaction

We use the 21-item Player Experience of Needs Satisfaction (PENS) scale [474] that measures the following dimensions:

1. Competence
2. Autonomy
3. Relatedness
4. Presence/Immersion
5. Intuitive Controls

The PENS instrument is based on self-determination theory (SDT) [116]. SDT asserts that there are three psychological needs that motivate individual behaviors that are universal, innate, and psychological: *competence* (seek to control outcomes and develop mastery [547]), *relatedness* (seek connections with others [30]), and *autonomy* (seek to be causal agents [94] while maintaining congruence with the self). SDT posits that these needs are vital to well-being. PENS contends that the psychological “pull” of games are largely due to their ability to engender these three needs [474]. PENS is consistently viewed as a robust framework for assessing player experiences [124, 458].

Game Experience Questionnaire

We use the 42-item Game Experience Questionnaire (GEQ) [114, 249, 250, 395, 435] that measures the following dimensions:

1. Flow, e.g., I felt completely absorbed.
2. Sensory and imaginative immersion, e.g., It was aesthetically pleasing.
3. Competence, e.g., I felt skillful.
4. Challenge, e.g., I felt challenged.
5. Tension, e.g., I felt frustrated.
6. Positive affect, e.g., I felt content.

7. Negative affect, e.g., I thought about other things.

The GEQ is both a widely used and recognized instrument [402].

Computer Programming Self-Efficacy Scale

Self-efficacy represents the belief in one's ability to succeed, either in a particular situation, or at a particular task [24]. The Computer Programming Self-Efficacy Scale (CPSES) is a scale for measuring programming self-efficacy. It consists of a validated 32-item scale that measures the following dimensions:

1. Independence and persistence
E.g., *Find ways of overcoming the problem if I got stuck at a point while working on a programming project.*
2. Complex programming tasks
E.g., *Organize and design my program in a modular manner.*
3. Self-regulation
E.g., *Find a way to concentrate on my program, even when there were many distractions around me.*
4. Simple programming tasks
E.g., *Write logically correct blocks of code using C++.*

Intrinsic Motivation Inventory

The Intrinsic Motivation Inventory (IMI) [367] assesses intrinsic motivation using four dimensions:

1. Interest/Enjoyment, e.g., I enjoyed doing this activity very much.
2. Effort/Importance, e.g., I put a lot of effort into this.
3. Pressure/Tension, e.g., I felt very tense while doing this activity.
4. Value/Usefulness, e.g., I believe this activity could be of some value to me.

Player Inventory Scale

The Player Inventory Scale (PIS) measures avatar identification [526], which consists of three second-order factors:

1. Similarity identification, e.g., My character is similar to me.
2. Embodied presence, e.g., In the game, it is as if I become one with my character.
3. Wishful identification, e.g., I would like to be more like my character.

Performance

Performance refers to the progress made in Mazzy as operationalized by the number of levels completed. Finer-grained distinctions are sometimes made, such as an analysis of the number of user attempts, number of hints, etc.

4.1.2 Recruitment

Participants are recruited via Amazon Mechanical Turk (AMT). AMT is a platform in which individuals can post Human Intelligence Tasks (HITs), e.g., marketing questionnaires, research studies, etc. Studies in a wide range of disciplines—social psychology, cognitive psychology, clinical psychology, etc.—have asserted that AMT provides data as good as more traditional methods, e.g., recruiting participants from college campuses [76, 352]. AMT workers tend to represent a more diverse sample than the U.S. population [41, 76, 88, 242, 255, 411, 514]. Many studies have shown AMT to be a reliable platform for experiments e.g., [76, 360].

Concerns Regarding Mechanical Turk

Researchers have argued that AMT participants may be “nonnaïve” as a result of having completed the same or very similar task before [89]. With a relatively large participant

pool spanning multiple studies, this is a valid concern. From the beginning, we carefully screened participants that had taken any previous study of ours—on the technical side, this was done by assigning each participant a flag that disqualified them from taking any future study of ours through a HIT requirement, i.e., each of our 10,000+ participants is a unique participant. Since our research environment was developed from scratch, there is minimal concern that they have experienced a similar HIT.

Other Precautions

We take a number of other precautions with Mechanical Turk workers:

- Limit HITs to workers with an acceptance rate > 90% or 95%
- Look for signs that a worker did not take the HIT seriously:
 - Bogus answers in free-text responses
 - Missing player data
 - Multiple surveys with zero variance between items
 - etc.

4.1.3 Design

We exclusively use a between-subjects experiment design (see Table 4.1). This means that participants are randomly assigned to one of a number of conditions. Methodologically, this is one of the most robust approaches to conducting experiments—counteracting carryover condition effects since participants are only exposed to one condition, and inter-subject differences since these will even out through random assignment.

A Note on Within-Subjects Designs

Within-subjects designs expose all participant to every condition. This is challenging when the conditions are substantially different avatars, and it may appear odd/suspicious

to the player were the avatar to change seemingly at random throughout the course of an experiment. Given the challenge with validity in the context of our theoretical questions, within-subjects designs are considered to be non-optimal.

4.1.4 Data Analyses

Data analyses generally involve the following:

- Statistical Analysis (t-test, ANOVA, MANOVA, chi-square test, etc.)
- Machine Learning (SVM, random forest, etc.)
- Other Computational (sentiment analysis, etc.)
- Other Qualitative (grounded theory, etc.)

4.2 Data Overview

The following is an aggregate overview of the data from the first 17 experiments listed in Table 4.1. These data exclude the badges study listed in Table 4.1. This overview also excludes studies not reported on in this thesis.

4.2.1 Participants Overview

See Figure 4-1 for a demographic profile of our participants. 4 out of the 17 studies, early pilot studies, did not collect demographic data, but are included for completeness. Race and gender data were collected in the same way as the current U.S. census.

Gender. We observe that 58% of participants identified as male and 42% female.

Race. We observe that 77% of participants identified as white, 6% as black or African American, 1% as American Indian or Alaska Native, 6% as South Asian, 2% as Chinese, 1% as Filipino, 0.4% as Japanese, 0.8% as Korean, 0.6% as Vietnamese, and 4.8% other.

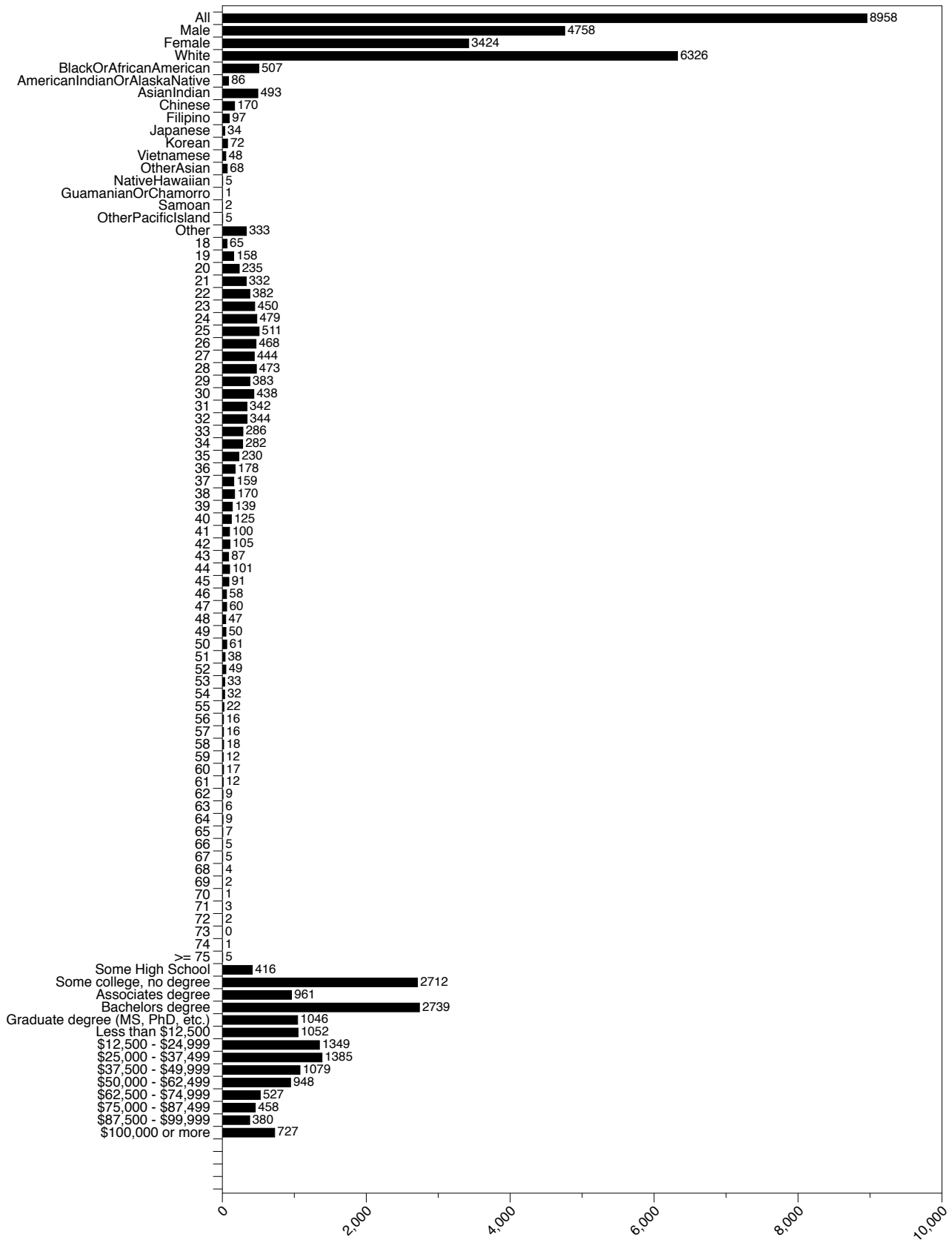


Figure 4-1: Demographic Profile

Age. We observe that the average participant’s age was 30.8 (SD = 9.8).

Education. We observe that most participants either had completed a bachelor’s degree (34.8%), or partial college (34.4%); while smaller numbers of participants held a graduate degree (master’s or doctorate) (13.3%), an associate’s degree (12.2%), or partial to full completion of high school (5%).

Household Income. We observe that most participants had a household income of \$25,000–\$37,499 (17.5%), or \$12,500–\$24,999 (17.1%). Smaller numbers of participants had a household income of \$37,500–\$49,999 (13.7%), < \$12,500 (13.3%), or \$50,000–\$62,499 (12.0%). Fewer participants had a household income of \$100,000 or more (9.2%), \$62,500–\$74,999 (6.7%), \$75,000–\$87,499 (5.8%), or \$87,500–\$99,999 (4.8%).

4.2.2 Summative Data Overview

The following is a table of *all* data described here: http://groups.csail.mit.edu/icelab/dkao/ALL_DATA.png. The table consists of *two sections*, a sample portion of the upper section is found in Figure 4-2, and a sample portion of the lower section is found in Figure 4-3. I now describe the two sections of the table in more detail.

Upper Section

The slice described here is from the *upper section* (Figure 4-2). Numbers in parentheses are *performance* and *engagement* (0-100), with the top number being the former and the bottom one below the asterisks being the latter. The *Version Mean* row is the mean across all participants in a particular game version (i.e., not discerning between avatar types). Roughly speaking, a performance > 50 means on average, participants completed > half the game, and engagement 50 means participants were neutral about the experience.








	# Participant	Game Versio	Protocol Vers	All	Male	Female
Per Avatar Type Effects (Performance/Engagement) Per Demographic						
Game Version #0						
	# Participant	Game Versio	Protocol Vers	All	Male	Female
Shape 	913	0	1	(51.56) ***** (57.01)	(56.66) ***** (56.52)	(43.65) ***** (55.69)
Likeness 	406	0	2	(43.34) ***** (53.53)	(33.29) ***** (47.62)	(36.78) ***** (54.98)
Minimal 	90	0	3	(46.47) ***** (59.6)	(###) ***** (###)	(###) ***** (###)
Friend 	104	0	2	(32.65) ***** (48.58)	(###) ***** (###)	(###) ***** (###)
Face Photo 	278	0	4	(50.47) ***** (54.07)	(56.31) ***** (49.79)	(47.83) ***** (57.36)
Scientist 	115	0	5	(46.71) ***** (67.21)	(44.94) ***** (68.24)	(49.37) ***** (65.8)
Scientist 	189	0	6	(53.07) ***** (54.43)	(50.97) ***** (55.11)	(55.53) ***** (53.6)
Version Mean				(46.32) ***** (56.35)	(48.43) ***** (55.46)	(46.63) ***** (57.49)

Figure 4-2: Example Data Slice; Upper








Per Avatar Type Effects (Performance/Engagement) Per Demographic							
Game Version #0		# Participant	Game Versio	Protocol Vers	All	Male	Female
					(3.02***) *****	(4.25) *****	(-2.37) *****
Shape		913	0	1	(0.91)	(1.15)	(-0.78)
Likeness		406	0	2	(-5.19**) ***** (-2.57*)	(-19.12) ***** (-7.75)	(-9.24) ***** (-1.5)
Minimal		90	0	3	(-2.06) ***** (3.5)	(###) ***** (###)	(###) ***** (###)
Friend		104	0	2	(-15.89***) ***** (-7.52*)	(###) ***** (###)	(###) ***** (###)
Face Photo		278	0	4	(1.94) ***** (-2.03)	(3.9) ***** (-5.58)	(1.81) ***** (0.88)
Scientist		115	0	5	(-1.83) ***** (11.11***)	(-7.48) ***** (12.87)	(3.35) ***** (9.32)
Scientist		189	0	6	(4.54) ***** (-1.67)	(-1.44) ***** (-0.26)	(9.51) ***** (-2.88)

Figure 4-3: Example Data Slice; Lower

Lower Section

The slice described here is from the *lower section* (Figure 4-3). Numbers in parentheses represent *deltas from the mean for all participants within a particular game version*. For example, choosing and using a shape avatar had an overall +3.02 effect on performance, and a +0.91 effect on engagement, as compared to *all* participants in *all* studies for *game version #0*². Therefore, we can read these numbers as relative to the entire demographic listed at the top. An asterisk by a number means this comparison is significant using a t-test (* p < .05, ** p < .01, *** p < .001).

²More specifically, the +/- is the mean of a specific row's participants minus the mean of participants in that game version (including that row's). The t-test also does a similar comparison, but to all other participants in the same game version (excluding that row's).

Supporting Information

Appendix [A](#) is a summary of game versions. Appendix [B](#) shows how aggregate performance and engagement statistics were calculated across all experiments. Appendix [C](#) is a summary of the experimental protocols.

Both experimental protocol and game version differ somewhat between avatar types due to iterative development. This includes other potential inter-experiment differences, such as time of testing during the calendar year. Although the table can be a good estimation at capturing the data, it is a rough guide only.

Overall Demographic Averages

The following figures summarize performance, engagement, and playtime averages over *all* demographics:

- **Performance:** [Figure 4-4](#).
- **Engagement:** [Figure 4-5](#).
- **Playtime:** [Figure 4-6](#).

Studying demographic differences is not the purpose of this work, but I do note there are a couple obvious trends. For instance, age appears correlated to engagement, and even more so with playtime, but age is inversely correlated to performance.

4.3 Experimental Trajectory

Overall, we find that avatar type significantly affects performance/engagement in across virtually all participants. These results constitute work establishing baseline understandings in an area of research that is underdeveloped. Our aim is motivated by a convergence of research in the social sciences establishing that identity plays an important role in learning.

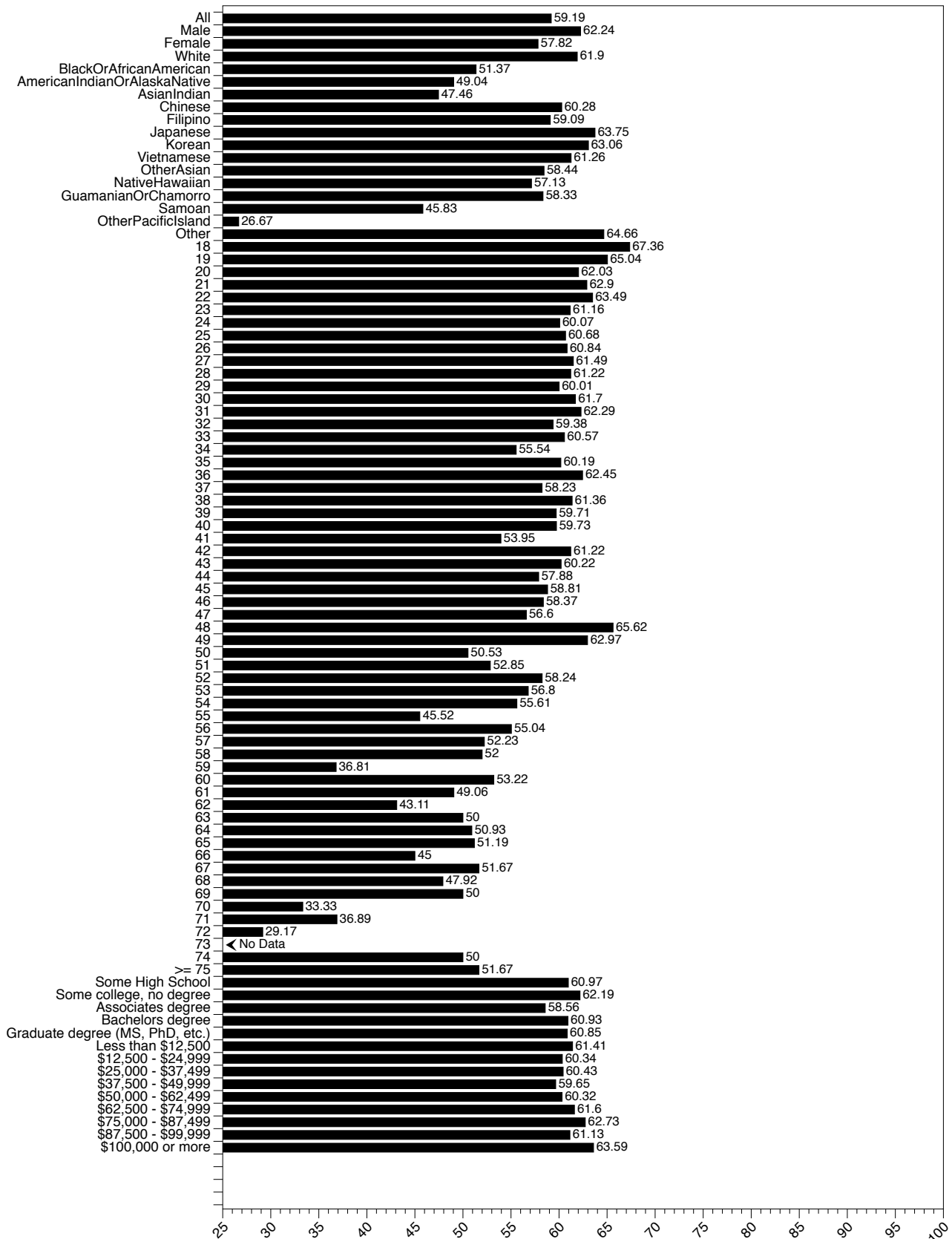


Figure 4-4: Performance Averages

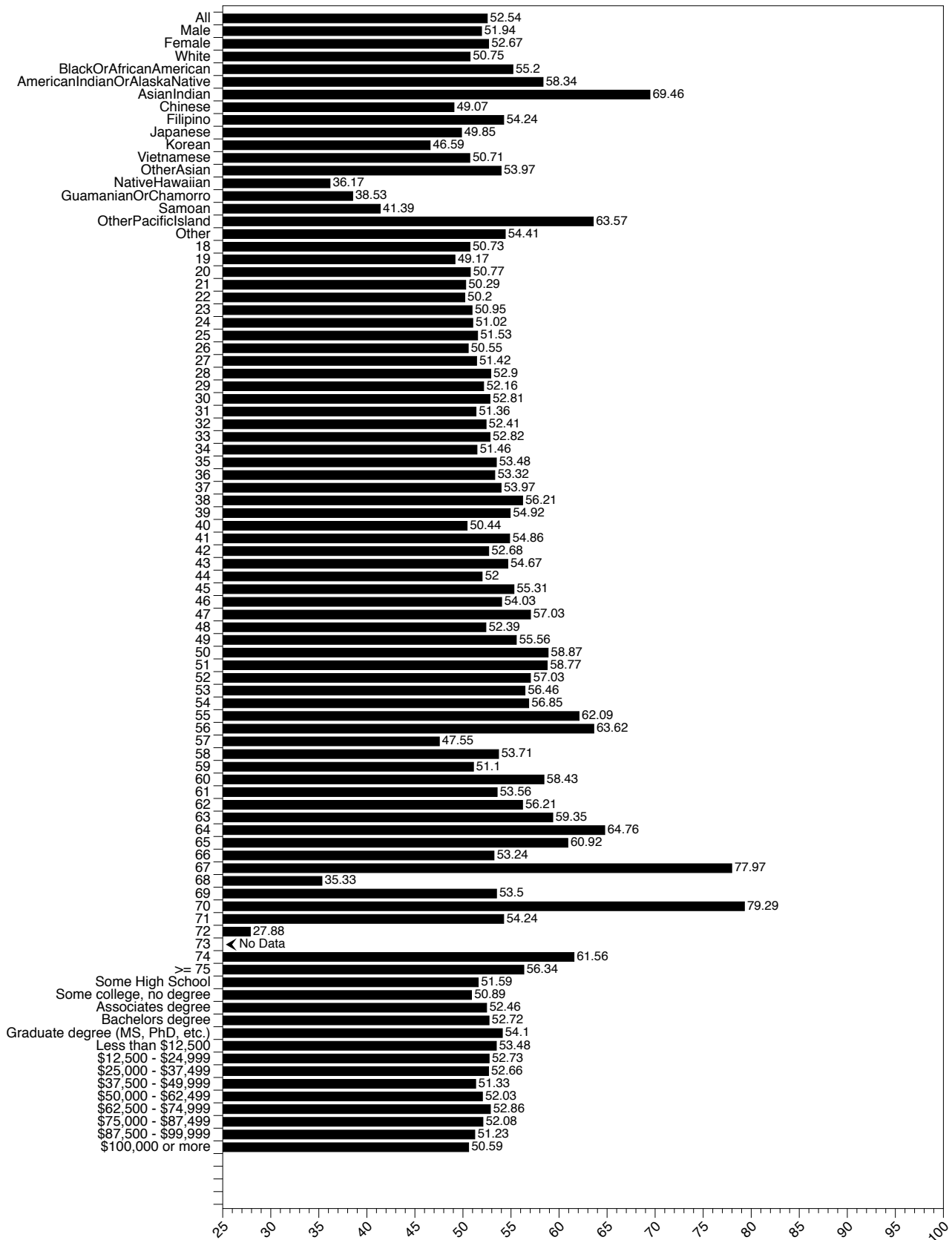


Figure 4-5: Engagement Averages

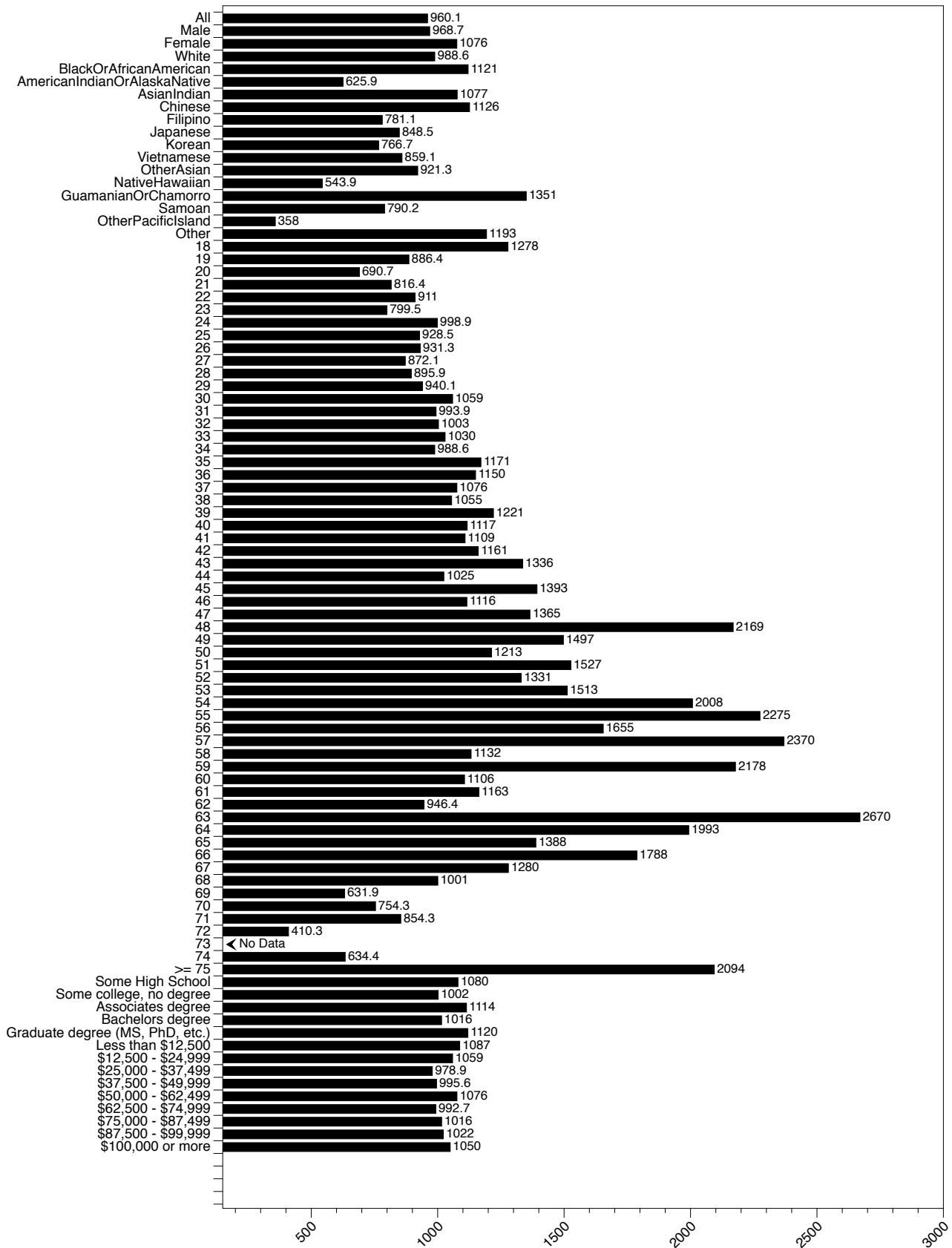


Figure 4-6: Playtime Averages

Our goal is not to say that one avatar is definitely better than another avatar—but rather to establish some baseline understandings regarding how avatars impact us differently.

4.3.1 Trajectory

As there are an infinite possible number of representations (e.g., animals, mythical creatures, people, famous people, your friends, furniture, indeed anything that can be used as your avatar representation)—while some assumptions can be made based on the social science literature, a lot of how these avatars impact us is simply unknown—it is necessary to explore varied representations.

Our experimental trajectory was initially based on understanding differences between anthropomorphic vs. non-anthropomorphic avatars, user likeness vs. non-likeness avatars, and investigating many sub-cases of each of these. For example, for non-anthropomorphic avatars we initially investigated various shape avatars, minimal (just a black dot) avatars, etc. For anthropomorphic we looked at Nintendo Miis that users customized in their own likeness, photos of users' actual faces, Miis customized in their best friend's likeness, etc. We then began to investigate famous scientists based on the social science literature suggesting that priming role models can improve academic performance. We expanded upon this by allowing users to Google image-search their own role models (e.g., Spiderman, famous actors, politicians, presidents, athletes, business magnates, musicians, etc.), or Google image-search their own shapes. We later refined this experiment into a larger-scale experiment comparing a pool of scientist avatars, to a pool of athlete avatars, to a pool of shape avatars. We then moved on to avatars that change during gameplay, since a large body of work suggested that identification facilitated vicarious outcomes, and detachment may facilitate outcome dissociation. Finally, we have studied other contextual and peripheral factors around avatars, such as presence of other players' avatars, the color of the avatar (in the case of a single color geometric shape), encouragement or lack thereof, winning vs. losing vs. nearly winning at an unrelated task prior to the experiment, and embellishment of background graphics.

4.3.2 Main Findings

Overall, we found that avatar type significantly affects performance/engagement in across virtually all participants. These results constitute work establishing baseline understandings in an area of research that is underdeveloped. Our aim is motivated by a convergence of research in the social sciences establishing that identity plays an important role in learning.

Here, I summarize a few of the notable findings from these experiments. This same summary is found at the end of the [Introduction](#). These findings represented what appeared to be the most consistent and powerful findings (i.e., multiple different studies confirmed it and/or it appeared to be a robust and convincing result). However, for the sake of conciseness, this subset necessarily leaves out other potentially interesting results, the full list of which can be found under each individual experiment in the next chapter.

Avatar-Based Outcomes:

- **Simple avatars often outperform complex avatars** [286]. This could be for a number of reasons. Seductive details [178], e.g., more complex, more embellished, etc. can be a distraction, outcome dissociation [286], e.g., non-human avatars promote less identification with failure, stereotype threat mitigation [503], e.g., simpler avatars contain fewer salient identity characteristics, and the Uncanny Valley, e.g., “almost” human avatars elicit revulsion [388].
- **Scientist role model avatars are extremely effective** [277, 279, 285]. Within a CS programming environment, all participants experience increased engagement while using scientist role model avatars, while female participants experience the most significant increases. Female participants often have significant increases in their play performance and reported engagement through using a well-known scientist as their avatar (e.g., Marie Curie), as compared to participants that used a well-known athlete as their avatar (e.g., Serena Williams), or a simple abstract shape (e.g., Triangle).
- **Successful likeness avatars can likely outperform any existing avatar types** [286]. We have discovered a new type of avatar, what we term the *successful likeness*. This

is a simple abstract avatar when the user is in the trial-and-error process and a likeness of the user only when the user achieves a goal. Compared to users that used only an avatar that was always simple abstract, or always a likeness of the user, or a likeness of the user when the user was in trial-and-error and a simple abstract avatar upon achieving a goal, these successful likeness participants played significantly longer and completed significantly more levels. We propose that these results can be explained by a model in which identification facilitates vicarious outcomes, and in which detachment facilitates outcome dissociation [286].

- **Red avatars cause significant decreases in engagement and avatar affect compared to blue avatars** [287]. Research has consistently shown that red reduces mood, affect and performance in cognitive-oriented tasks [146, 190, 244, 271, 314, 329, 374, 376, 493]. For example, Lichtenfeld et. al showed that even just peripherally noticing red (e.g., hidden in a question, in the copyright notes at the end of a page, etc.) can have similar effects [329]. Prior work on first-person shooter (FPS) multiplayer games have hypothesized that blue teams are at a disadvantage because they “see red” [251]. We provide the first study to show that this effect is true in a single-player context [287]. This red-blue discrepancy was higher for male players than for female players.
- **Badges and avatar identification promote positive outcomes** [290]. We have found that badges can promote avatar identification (personal interest, role model), player experience (achievement, role model), intrinsic motivation (achievement, role model), and programming self-efficacy (role model) during both game play and game making. Independently of badges, avatar identification promotes player experience, intrinsic motivation, programming self-efficacy, and the total time spent playing and making. Avatar identification also promoted other meaningful in-editor activity, such as playtesting time, etc. and led to significantly higher overall quality of the completed game levels (as rated by 3 independent externally trained QA testers) [290].

Other Outcomes:

- **Positive and neutral encouragement text displayed at regular intervals (e.g., “Keep it up!”), significantly increases engagement as compared to no text or neg-**

ative encouragement text [288]. Encouragement is different from feedback, in that it does not necessarily encode information about performance [303, 384, 444, 478]. Regularly dispensed encouragement, operationalized as text appearing at the bottom of the screen—both positive (e.g., “You’re doing good”) and neutral (e.g., “You’re doing average”) significantly increased player engagement as compared to negative (e.g., “You’re doing badly”) or none.

- **More embellished game backgrounds cause players to have significantly decreased game performance and significantly decreased programming self-efficacy but significantly increased engagement** [289]. Research suggests that the addition of seductive visual details in video games hinders performance of learners [178, 455, 513]. Yet, other research results propose the opposite: that visual embellishments and well-designed ambiguity instead improve learners’ performance, engagement, and self-efficacy [488, 517, 554]. To shed light on this apparent contradiction, we implemented the following four game themes: 1) *Generic* theme with no embellishments (simple flat color background), 2) *Fantasy* game theme (forest, snow, and desert adventure backgrounds), 3) *STEM-oriented* theme (computer circuitry background), and 4) *Choice* (the user picks one of the previous three options). Generic condition participants had highest performance (levels) and had highest programming self-efficacy—followed by choice, fantasy game setting, circuitry. However, ordering of conditions for engagement was precisely opposite the trend for performance. These are trade-offs between two diametrically opposed approaches to game themes and embellishment: instrumental game skins vs. thematic and deliberately embellished game skins [289].

Chapter 5

Experiments

This chapter describes each experiment (see Table 5.1) in more detail. Experiments are grouped under four different headings: [Central Avatar Experiments](#), [Additional Avatar Experiments](#), [Interface Experiments](#), and [Culminating Experiment](#).

Central Avatar Experiments are primary to this dissertation, and include the main avatar studies performed. Because an avatar can look like anything, we had to first look for some basic distinctions to begin to understand their impacts. The very first such distinction we made was between anthropomorphic and non-anthropomorphic avatars [389, 539]. As such, we started by studying likeness avatars versus shape avatars ([Shape vs. Likeness #1/#2](#)), likeness avatars versus minimalistic avatars ([NoAvatar vs. Likeness](#)), and friend likeness avatars versus shape avatars ([Shape vs. Friend](#)). We also made a distinction between photo-realistic and non-photo-realistic likenesses ([Likeness vs. EasyLikeness](#)), with the photo-realistic likenesses also being compared to shape avatars ([Shape vs. InstantLikeness](#)). These studies were done to explore a range of both anthropomorphic and non-anthropomorphic avatars. We found results from the preceding studies that suggested stereotype threat could be a potential factor. As a result, we began to study role models as avatars ([ScientistText vs. ShapeText](#), [Shape vs. Scientist](#), [Shape vs. RoleModel](#), and [Shape vs. Scientist vs. Athlete](#)). Finally, the literature has suggested that we live vicariously through our avatars, as such, we explored a new type of avatar that is a likeness *during success only* ([Successful Likeness](#)).

Additional Avatar Experiments explores studies related to avatars that are less central to this thesis. We studied how the presence of other player avatars affected users as a type of worked example and competition/collaboration mechanic ([Phantoms vs. Non-Phantoms](#)) and specifically how red avatars and blue avatars differed in their impact on users ([Red vs. Blue](#)). We performed these studies since many digital environments feature multiple users, and because the color alone of avatars is a low-level feature of *every* avatar.

Interface Experiments sought to study other contextual factors at the interface level of our platform. While our initial goal was to understand the potential interaction effects of factors like style and genre, we realized that the richness of these questions deserved investigation in their own right. We studied encouragement ([Feedback Positive vs. Negative vs. Neutral vs. Nothing](#)), nearly-winning at an unrelated task prior to engaging in our platform ([Mini-Game Loss vs. Near-Win vs. Win](#)), and visual themes and embellishment ([Game Theme Basic vs. Circuit vs. RPG vs. Choice](#) and [Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice](#)). We performed these studies since these types of questions are even more broadly relevant to digital systems in general.

Culminating Experiment is a culmination of previous studies. We study how role model *badges* can be effective, while using a standard likeness avatar. In addition to role model badges, we also look at achievement badges and personal interest badges. Separately from badges, we also study avatar identification and whether it is able to predict various game and game-making outcomes ([Culminating Experiment](#)). We performed this study to better understand avatar identification and whether badging can be an alternative form of expressing avatar types.

I give a brief overview only for some experiments that are either very similar to one we conducted already, or that are highly exploratory.

Experiment	N
Shape vs. Likeness #1	258
Shape vs. Likeness #2	250
NoAvatar vs. Likeness	182
Shape vs. Friend	208
Likeness vs. EasyLikeness	128
ScientistText vs. ShapeText	224
Shape vs. Scientist	399
Shape vs. InstantLikeness	446
Shape vs. RoleModel	357
Shape vs. Scientist vs. Athlete	1067
Phantoms vs. Non-Phantoms	523
Successful Likeness	997
Red vs. Blue	507
Feedback Positive vs. Negative vs. Neutral vs. Nothing	645
Mini-Game Loss vs. Near-Win vs. Win	366
Game Theme Basic vs. Circuit vs. RPG vs. Choice	1171
Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice	1230
Badge Type Comparison; 6 Conditions	2189

Table 5.1: Experiment Summary from Intro.

5.1 Central Avatar Experiments

Experiment listing:

[Shape vs. Likeness #1/#2](#)

[NoAvatar vs. Likeness](#)

[Shape vs. Friend](#)

[Likeness vs. EasyLikeness](#)

[ScientistText vs. ShapeText](#)

[Shape vs. Scientist](#)

[Shape vs. InstantLikeness](#)

[Shape vs. RoleModel](#)

[Shape vs. Scientist vs. Athlete](#)

[Successful Likeness](#)

5.1.1 Shape vs. Likeness #1/#2

Category: [Central Avatar Experiments](#)

Next Experiment: [NoAvatar vs. Likeness](#)

Experiment Overview (Shape vs. Likeness #1/#2)

This experiment presents results of a comparative study between avatars in the likeness of players and avatars as geometric shapes. In our STEM learning game, results show that players that had selected and used a shape avatar had significantly higher performance than players that had customized and used a likeness avatar. Players using the shape avatar also had significantly higher self-reported engagement, despite having lower self-reported affect towards the avatar.

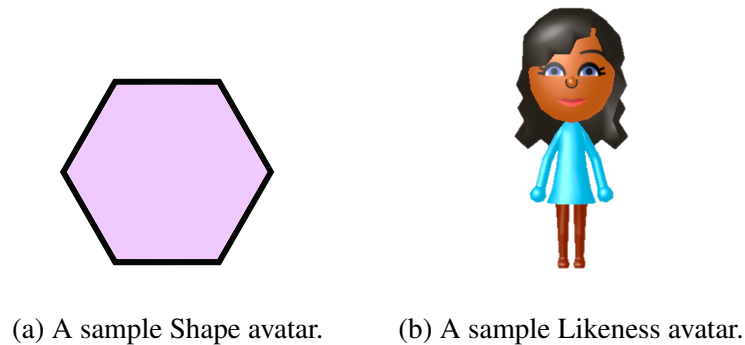


Figure 5-1: Sample avatars.

Experiment-Specific Methods (Shape vs. Likeness #1/#2)

The study we performed consisted of two experiments (N=508) inside of our educational game *Mazzy*. The study compares the impact of selected shape avatars and customized likeness avatars on player engagement and performance.

Avatar Conditions

The two avatar conditions we tested were:

1. Shape: Avatar as a geometric shape.
2. Likeness: Avatar in the likeness of the player.s

The shape condition was a geometric shape; players picked out of eight possible geometric shapes [362]. These players were told the shape that they picked would be their avatar in the ensuing game. The likeness condition consisted of a Mii avatar. A Mii is a character developed by Nintendo, chosen since Miis were designed with the intention of looking similar to users (Mii is a blend of “Wii” and “me”). Players were asked to use a publicly available Mii customization system prior to the task [14]. Furthermore, players were told to create an avatar that looked like themselves and that this avatar would be used in the subsequent game. See Figure 5-1 for examples of these.

Task

The experimental task was to play the first version of *Mazzy*. See Section 3.1.6 and the subheading [Final Build of First Version](#) for a description.

Quantitative and Qualitative Measures

The performance measures we recorded were:

- **Levels completed:** The number of levels completed.
- **Level attempts:** The number of attempts in each level.
- **Level bonus items:** Bonus items collected in each level.

The engagement measures we recorded were:

- **Enjoyment:** Enjoyment rating in each level.
- **Difficulty:** Difficulty rating in each level.

All subjective data were collected using a 5-point Likert scale. These engagement measures (enjoyment and difficulty) were the only engagement data collected in the first experiment. In the second experiment, at the end of the study, players were also asked to rate how they felt overall with respect to the game, their progress, and their avatar, in addition to describing their avatar in text and completing a demographics survey.

Participants

508 participants (250 in the first experiment, 258 in the second experiment) were recruited through Mechanical Turk. 38% of the participants were female. 77% of participants were white, 9% black or African American, 5% Chinese, the remaining participants were divided amongst eleven other group categories. Participants were between the ages of 18-68 ($M = 31.6$) and were reimbursed \$2 to participate.

Design

A between-subjects design was used: avatar type was the between-subjects factor. Participants were randomly assigned to conditions (i.e., random assignment of avatar type).

Experiment Protocol

Prior to starting the task, players were told they could exit the game *at any time*. Then, for each condition players loaded the game in their web browser. After each level that players completed, players were presented with a screen showing the number of “stars” they had earned (corresponding to the number of bonus items they had collected); at this point in the procedure, players could either continue or replay the level. If they chose to replay the level, they were brought back to the previous level (with their previous code still intact). If they continued, they were then asked to report engagement (enjoyment and difficulty). When participants were done playing, they returned to the instructions, which prompted them with additional questions including the demographic survey.

Analysis

Our analysis consists of independent-samples t-tests, and results are reported as significant when $p < 0.05$ (two-tailed). Furthermore, we perform linguistic analysis and supervised learning as described below.

Natural Language Processing

The text we want to analyze are players’ linguistic descriptions of their avatars, typically 2-3 sentences long. In order to interpret these, we leverage a text analysis system called Linguistic Inquiry Word Count (LIWC). LIWC is a popular tool in psychology. LIWC was developed over the last couple decades by human judges that categorize common words [427, 429]. LIWC matches text to 82 language dimensions; these range from affective processes (i.e., positive emotion, negative emotion, anxiety, etc.) to part-of-speech (i.e., articles, past tense, present tense, etc.) to thematic categories (i.e., achievement, money, death, etc.). Pennebaker et. al performed one of the earliest text analyses, using sources such as daily diaries and journal abstracts; they found that linguistic style is a “meaningful way of exploring personality” [428]. In our case, we leverage LIWC to analyze players’ descriptions of their game avatars; we are interested in exploring how players perceive themselves in relation to their avatars. We use LIWC to calculate scores for each player

individually, then present the averages for each condition in the results. Given the large number of language dimensions analyzed by LIWC, we present only results from ten of the dimensions that have the highest difference in score between avatar types.

Prediction Algorithms

Here, we are looking to test the effectiveness of a player model that incorporates social and virtual identity in predicting when players will quit our game. In order to make these predictions, we must select some subset of machine learning algorithms to train and test on. We use the WEKA machine learning workbench (version 3.7.12). WEKA was developed at the University of Waikato [207], and contains a collection of machine learning algorithms for data mining tasks. This version of WEKA has by default over 50 different classification algorithms. Furthermore, WEKA's package manager gives access to an additional set of classification algorithms; this makes the total number of classification algorithms available close to 100. Given the large number of choices, we use a similar approach to Mahlmann et. al in that we consider at least one algorithm from each of the families of algorithms [348]. Similarly, we pay especially close attention to algorithms found on the list of top ten data mining algorithms: SVMs, decision trees, belief networks, etc. [559]. The specific attributes used in the algorithms is as follows:

- **Avatar Type:** The avatar type (Likeness, or Shape).
- **Avatar Shape:** The avatar sub-type. For likeness avatars, these were coded as "Mii"; for shape avatars, these were coded as "Triangle", "Square", "Pentagon", etc.
- **Level One Enjoyment:** Player reported enjoyment in level 1.
- **Level One Difficulty:** Player reported difficulty in level 1.
- **Level One Stars:** Number of bonus items in level 1.
- **Level One Attempts:** Number of attempts in level 1.
- **Level One Successful Attempts:** Number of succ. attempts in level 1.
- **Player Age:** The player's age.
- **Player Gender:** The player's gender.
- **Player Race:** The player's race.

Table 5.2: Results from the first experiment.

Attribute	L-Mean	L-SD	S-Mean	S-SD	t-test
Levels Completed	1.90	1.14	1.96	1.16	0.43
Average Enjoyment	2.89	1.11	3.26	0.96	2.35*
Average Difficulty	2.34	0.95	2.32	0.86	0.26
Total Bonus Items	3.10	3.28	3.18	3.38	0.19
Total Attempts	21.87	21.60	18.65	15.21	1.40

* $<.05$, ** $<.01$, L = Likeness, S = Shape, SD = Standard Deviation

We used a simple single-attribute evaluator called 1R to rank these attributes by importance. 1R generates a one-level decision tree that splits on a single attribute (i.e., all predictions for that tree depend only on that specific attribute). 1R has been shown to perform well vis-à-vis more complex algorithms [239]. We use the 1R evaluator on each attribute individually; we then rank those attributes by their prediction scores, giving us a rough approximation of each attribute's merit.

Results & Findings (Shape vs. Likeness #1/#2)

Experiment 1

Players reported higher engagement in the shape condition. Players in the shape condition ($M=3.26$, $SD=0.96$) reported significantly higher enjoyment than participants in the likeness condition ($M=2.89$, $SD=1.11$), $t(205)=2.35$, $p=0.02$. No other significant differences were found. See Table 5.2.

Experiment 2

Players had higher performance and engagement in the shape condition. Players had lower affect towards the shape avatar. Players in the shape condition ($M=1.65$, $SD=1.07$) completed significantly more levels than participants in the likeness condition ($M=1.08$, $SD=1.01$), $t(256)=4.42$, $p=0.0001$. As a result, players in the shape condition ($M=16.14$, $SD=12.86$) had more total attempts than participants in the likeness condition ($M=12.38$,

Table 5.3: Results from the second experiment.

Attribute	<i>L</i> -Mean	<i>L</i> -SD	<i>S</i> -Mean	<i>S</i> -SD	t-test
Levels Completed	1.08	1.01	1.65	1.07	4.42**
Average Enjoyment	2.86	0.88	3.05	0.95	1.44
Average Difficulty	2.15	0.82	2.23	0.88	0.62
Total Bonus Items	1.99	2.73	2.69	3.08	1.93
Total Attempts	12.38	10.99	16.14	12.86	2.52*
Avatar Rating	3.61	0.94	3.06	0.87	4.84**
Progress Rating	3.06	1.08	3.34	1.03	2.09*
Game Rating	3.07	1.10	3.45	1.02	2.89**

* $<.05$, ** $<.01$, L = Likeness, S = Shape, SD = Standard Deviation

SD=10.99), $t(255)=2.52$, $p=0.01$. Players in the shape condition ($M=3.45$, $SD=1.02$) rated the game higher than participants in the likeness condition ($M=3.07$, $SD=1.10$), $t(253)=2.89$, $p=0.004$. Players in the shape condition ($M=3.34$, $SD=1.03$) also rated their progress higher than participants in the likeness condition ($M=3.06$, $SD=1.08$), $t(252)=2.09$, $p=0.038$. Players in the likeness condition ($M=3.61$, $SD=0.94$) rated their avatar higher than participants in the shape condition ($M=3.06$, $SD=0.87$), $t(254)=4.84$, $p=0.0001$. Overall trends remain consistent across both experiments. See Table 5.3.

Text Analysis

Table 5.4 contains a summary of text analysis results on players' descriptions of their avatars. Figures 5-2 and 5-3 are word clouds of players' avatar descriptions. Common english words, as well as the words "avatar" and "game" have been removed from these clouds to highlight differences.

Level Prediction

To determine the usefulness of modeling aspects of social and virtual identities, we built a player model using only statistics from the first level. We then ran a number of machine learning algorithms to determine if we could predict the final level completed. This involved removing those participants that did not complete the first level (there were 73 such partici-

Table 5.6: Prediction accuracy of various machine learning algorithms. Higher means that the algorithm performed better.

Algorithm	Accuracy
C4.5	51.9%
LibSVM	51.3%
Random Forest	47.6%
Bayes Network	47.0%
Multilayer Perceptron	45.4%
k-Nearest Neighbors	42.7%
Logistic Regression	42.7%
Baseline	40.0%

Table 5.7: The 1R attribute evaluation scores for each feature.

Attribute	1R Score
Avatar Type	47.03
Avatar Shape	44.32
Level One Attempts	43.24
Player Age	43.24
Player Race	41.08
Level One Succ. Attempts	41.08
Level One Difficulty	39.46
Player Gender	36.76
Level One Enjoyment	36.76
Level One Stars	34.60

items collected, reported engagement (enjoyment and difficulty), and gender were the least effective individual predictors.

Experiment-Specific Discussion (Shape vs. Likeness #1/#2)

The results suggest that avatar type has a significant impact on user performance and engagement in our STEM learning game. This has important implications. Level completion in an educational game can be seen as evidence that learning has occurred (so long as the

task is novel). Therefore, understanding virtual identities' impacts may be crucial in better understanding how they affect learners in educational games.

We might ask why specifically there was a large, measurable difference in performance and engagement between these two avatar types. Depending on the point of view one takes, this can be explained by a number of phenomena. Bowman et. al suggest that avatars more like "objects" cause players to focus more on in-game mechanics and challenges ("pleasures of control") [63]. There is evidence of this in the text analysis. Players detailing their shape avatars are more likely to use impersonal pronouns (e.g., it, it's, those) and articles (e.g., a, an, the), and less likely to use first person singular (e.g., I, me, my). Failure in the game (which is almost guaranteed, the mean number of attempts in the first level was 8.4), may be especially thwarting when the character failing is *you*. This would suggest that, for instance, failing as an abstract shape, but succeeding as a likeness to yourself, would be an effective adaptive avatar representation for learning.

Players using likeness avatars often made personal comparisons, e.g., "My avatar has a likeness to myself [...] she is chubby like me.", "[...] I had black hair and a gray shirt and my red glasses", and one player commented "[...] the avatar's success is my own", seeming to support the above. But some players felt they were unable to adequately represent themselves in the Mii; one participant said "it was difficult to make the avatar look like me" and "there weren't enough colors to customize the shirts." This means that despite the large number of options for hairstyles (72), eyes (48), mouths (24), etc. some players still found the avatar creator to be limiting. Even though this is the case, we found the avatar creator to be more than sufficient for most players. Figure 5-4 suggests that stereotype threat may have been an additional contributing factor for some players; there were greater disparities between the two avatar types in African American players, i.e., the likeness avatar may have acted as a stereotype threat trigger, as consistent with previous work [283].

Because the actual customization of the likeness avatar was part of the condition, perhaps players simply did not enjoy that aspect of the game. Or it is possible they did enjoy it, but were unsatisfied with the avatar's role in the game. If this was the case, it affected not

only their performance, but also significantly affected their disposition towards the game in a negative manner, *despite* the fact that player avatar ratings are strongly in favor of the likeness avatar. It is clear that more work needs to be done in distinguishing the specific psychological effects at play here. However, the results suggest that there are differences between avatars customized in the *likeness of players* and avatars selected as *geometric shapes*. Were we to make a recommendation to educational game makers based on these results alone, we would be hard-pressed to make a definite statement. However, if faced between simpler, abstract avatars and more complex, customizable avatars, we would be in support of simpler avatars.

5.1.2 NoAvatar vs. Likeness

Previous Experiment: [Shape vs. Likeness #1/#2](#)

Category: [Central Avatar Experiments](#)

Next Experiment: [Shape vs. Friend](#)

Experiment Overview (NoAvatar vs. Likeness)

In this exploratory study, we compared an avatar that consisted only of a black dot (NoAvatar) vs. a Mii likeness (Likeness) avatar. We did this to see if the absence of shape selection would impact the results. Conducted similarly to Shape vs. Likeness #1/#2, we found similar results with the NoAvatar condition outperforming the Likeness condition.

5.1.3 Shape vs. Friend

Previous Experiment: [NoAvatar vs. Likeness](#)

Category: [Central Avatar Experiments](#)

Next Experiment: [Likeness vs. EasyLikeness](#)

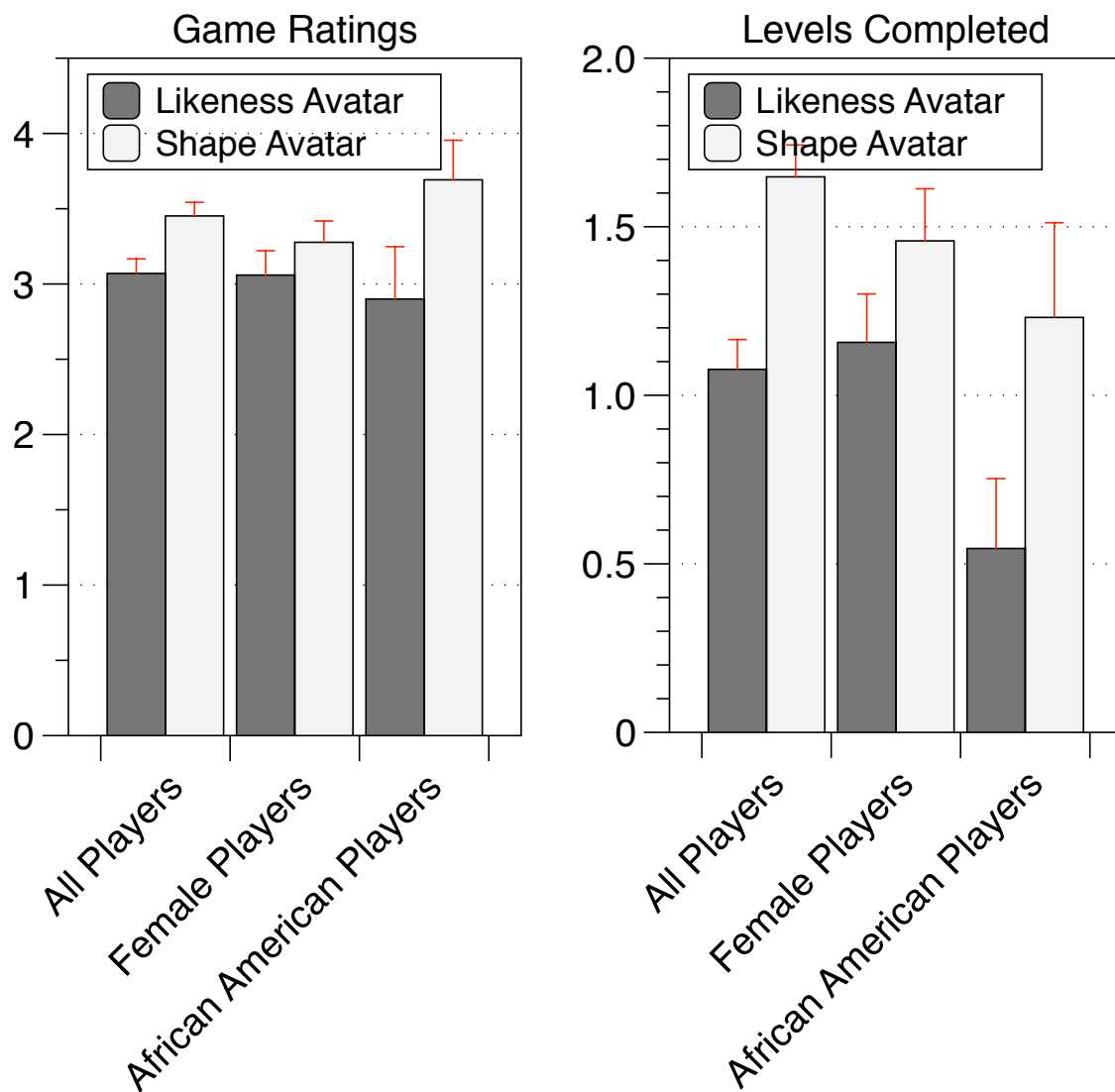


Figure 5-4: Experiment 2 game ratings and level completion averages between avatar types across social categories. Here, the focus is on two social groups underrepresented in STEM (95% CIs).

Experiment Overview (Shape vs. Friend)

In this exploratory study, we compared an avatar that was a shape vs. a Mii customized to look like one of the user's friends (Friend). We did this to see if making an avatar in the likeness of a person other than the self would impact the results. Conducted similarly to Shape vs. Likeness #1/#2, we found similar results with the Shape condition outperforming the Friend condition.

5.1.4 Likeness vs. EasyLikeness

Previous Experiment: [Shape vs. Friend](#)

Category: [Central Avatar Experiments](#)

Next Experiment: [ScientistText vs. ShapeText](#)

Experiment Overview (Likeness vs. EasyLikeness)

In this exploratory study, we compared an avatar that was a Mii likeness (Likeness) vs. a photo of the user (EasyLikeness). We did this to see if customizing an avatar to look like oneself would lead to a different result than using a photo. Conducted similarly to Shape vs. Likeness #1/#2, we found no significant differences.

5.1.5 ScientistText vs. ShapeText

Previous Experiment: [Likeness vs. EasyLikeness](#)

Category: [Central Avatar Experiments](#)

Next Experiment: [Shape vs. Scientist](#)

Experiment Overview (ScientistText vs. ShapeText)

In this exploratory study, we compared an avatar that was a scientist (ScientistText) vs. a shape (ShapeText). Snippets from Wikipedia were included alongside each avatar. See Protocol #5 and #1 in [Protocol Versions](#). We did this to see if scientist avatars would affect participants positively. As we found a trend of higher female participant performance in the scientist condition, this was a direct motivation for our following study.

5.1.6 Shape vs. Scientist

Previous Experiment: [ScientistText vs. ShapeText](#)

Category: [Central Avatar Experiments](#)

Next Experiment: [Shape vs. InstantLikeness](#)

Experiment Overview (Shape vs. Scientist)

In this exploratory study, we compared an avatar that was a shape vs. a scientist. Conducted similarly to ScientistText vs. ShapeText but with minor differences (no avatar text snippets and a different set of scientists—Protocol #6 in [Protocol Versions](#)), we found the following.

Results & Findings (Shape vs. Scientist)

Female participants in the scientist condition outperformed female participants in the shape condition. Female participants in the scientist condition completed significantly more levels ($M=1.7$, $SD=1.1$) than female participants in the shape condition ($M=1.3$, $SD=1.0$), $t(179)=2.51$, $p=0.006$, $d=0.38$. Female participants in the scientist condition collected significantly more bonus items ($M=2.9$, $SD=3.2$) than female participants in the shape condition ($M=1.4$, $SD=2.4$), $t(179)=3.62$, $p=0.0002$, $d=0.54$. A chi-square test found that female participants chose different scientists than male participants ($X^2=27.8$, $p=0.0002$,

$V=0.38$), with the female participants most preferring Marie Curie (35%), and the male participants most preferring Albert Einstein (39%).

5.1.7 Shape vs. InstantLikeness

Previous Experiment: [Shape vs. Scientist](#)

Category: [Central Avatar Experiments](#)

Next Experiment: [Shape vs. RoleModel](#)

Experiment Overview (Shape vs. InstantLikeness)

In this exploratory study, we compared an avatar that was a shape vs. a photo of the user (InstantLikeness). We did this to see if using a photo likeness of oneself would lead to different results than a customized avatar. Conducted similarly to Shape vs. Likeness #1/#2, we found that African American participants had decreased positive affect in the InstantLikeness condition as opposed to all other participants who had similar positive affect between InstantLikeness and Shape. All participants reported higher level difficulty in the InstantLikeness condition [283].

5.1.8 Shape vs. RoleModel

Previous Experiment: [Shape vs. InstantLikeness](#)

Category: [Central Avatar Experiments](#)

Next Experiment: [Shape vs. Scientist vs. Athlete](#)

Experiment Overview (Shape vs. RoleModel)

Research has indicated that role models have the potential to boost academic performance [356, 357]. In this experiment, we explore role models as game avatars in an educational



Figure 5-5: Player selected role model avatars.

game. Of particular interest are the effects of these avatars on players' performance and engagement. Participants were randomly assigned to a condition: a) user selected role model avatar, or b) user selected shape avatar. Results suggest that role models are heavily preferred. African American participants had higher game affect in the role model condition. South Asian participants had higher self-reported engagement in the role model condition. Participants that completed ≤ 1 levels had higher performance in the role model condition. General trends suggest that the role model's gender and racial closeness with the player, could play a role in player performance and self-reported engagement as consistent with the social science literature.

Experiment-Specific Methods (Shape vs. RoleModel)

The study we performed consisted of an experiment (N=357) inside of the game Mazzy. The study compares the impact of player selected role model avatars versus player selected shape avatars on player engagement and performance.

Avatar Conditions

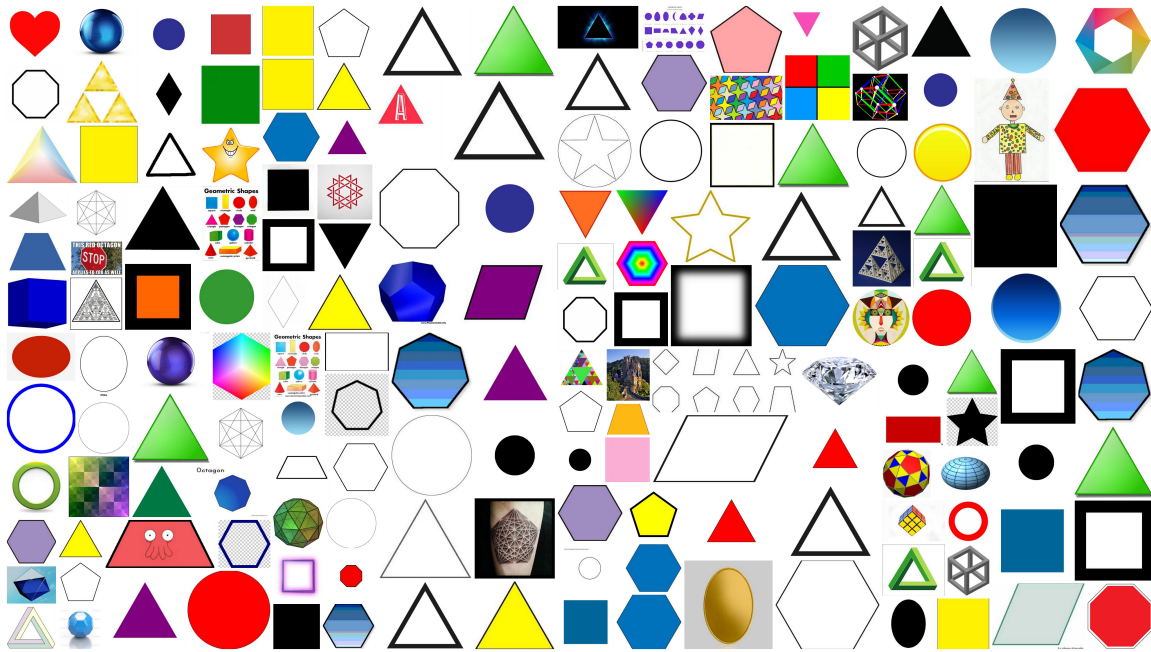


Figure 5-6: Player selected geometric shape avatars.

The two avatar conditions we tested were:

1. Role Model: Avatar in the form of a role model.
2. Shape: Avatar in the form of a geometric shape.

Role model condition participants were asked to think of any type of role model (real or fictional). Shape participants were asked to think of any type of geometric shape. Participants were informed that whatever they came up with would be their game avatar. Participants then used Google image search to find an image representing their choice. This image was uploaded to the game and became the user's character that moved about the maze.

Task

The experimental task was to play the first version of *Mazzy*. See Section 3.1.6 and the subheading [Final Build of First Version](#) for a description.

Quantitative and Qualitative Measures

The performance measures we recorded were:

- **Levels completed:** The number of levels completed.
- **Level attempts:** The number of attempts in each level.
- **Level bonus items:** Bonus items collected in each level.

The engagement measures we recorded were:

- **Enjoyment:** Enjoyment rating in each level.
- **Difficulty:** Difficulty rating in each level.

All subjective data was collected using a 5-point Likert scale. Players were also asked at the end of the experiment to rate how they felt overall with respect to the game, their progress, and their avatar, in addition to completing a demographics survey.

Participants

357 participants were recruited through Mechanical Turk. 129 of the participants were female. 193 of the participants were white, 112 south Asian, 22 black or African American, and the remaining participants divided among eleven other group categories. Participants were between the ages of 19-65 ($M = 31.4$) and were reimbursed \$2 to participate.

Design

A between-subjects design was used: avatar type was the between-subjects factor. Participants were randomly assigned to conditions.

Experiment Protocol

Prior to starting the task, players were told they could exit the game at any time. Then, for each condition players loaded the game in their web browser. After each level that players completed, players were presented with a screen showing the number of “stars” they had earned (corresponding to the number of bonus items they had collected); at this point in the procedure, players could either continue or replay the level. If they chose to replay the level, they were brought back to the previous level (with their previous code still intact). If they continued, they were then asked to report engagement (enjoyment and difficulty). When participants were done playing, they returned to the instructions, which prompted them with

additional questions including the demographic survey.

Analysis

Data was extracted and imported into Statistical Package for Social Science (SPSS) version 22 for data analysis using multivariate analysis of variance (MANOVA). The dependent variables are—*total levels completed, total attempts, total bonus items, average enjoyment, average difficulty, avatar rating, progress rating, game rating*; and the independent variables are—*avatar type (role model vs. shape), gender, race*. All the dependent variables are continuous variables. For the independent variables, both the avatar status (i.e., 0 = shape; 1 = role model) and gender (i.e., 0 = female; 1 = male) were dichotomous variables, and race (i.e., 1 = white, 2 = black or African American, 3 = south Asian, 4 = other) is a categorical variable. To detect the significant differences between user role model and user shape avatars, we utilized two-way or factorial MANOVA. The reason we chose factorial MANOVA is that we suspected that there would be an interaction effect between the independent variables. Also, we considered the variable—*age* as a covariate in the analysis (using MANCOVA), however, age was found not to be a significant covariate, as a result, it was not included in the subsequent analyses. First, we ran two-way MANOVA with avatar type and gender as independent variables, and then, another two-way MANOVA with avatar type and race. We also ran targeted independent-samples t-tests on the following groupings: *low performers (completed ≤ 1 levels), high performers (completed ≥ 2 levels), same vs. different gender role models, and same vs. different race role models*. These results are reported as significant when $p < 0.05$ (two-tailed). Before running MANOVAs, all the variables included in the analyses were checked, and there were no outlier detected. Prior to running our first MANOVA model, we checked both assumption of homogeneity of variance and homogeneity of covariance by the test of Levene's Test of Equality of Error Variances and Box's Test of Equality of Covariance Matrices; and both assumptions were met by the data ($p > .05$ for Levene's Test, and $p > .05$ for Box's Test). And for our second MANOVA model, using same tests, both assumptions were found not tenable ($p < .05$ for Levene's Test except Average Enjoyment Rating and Levels completed, and $p < .05$ for Box's Test). To address this violation issue, Pillai's Trace value was considered instead of Wilk's Lambda

value.

Results & Findings (Shape vs. RoleModel)

Avatar and Gender

Participants in the user role model condition had significantly higher avatar ratings compared to participants in the user shape condition. Our first MANOVA model that contained avatar types and gender as independent variables with a set of eight dependent variables was tested first. In this model, we looked for the main effect of avatar types, another main effect of gender, and an interaction effect of avatar types as well as gender on the set of dependent variables. The test results of the first MANOVA model indicated the main effect of avatar types as significant ($\lambda = .784$, $F(8, 244) = 8.399$, $p < .001$) whereas the main effect of gender and the interaction effect of avatar and gender were found not be significant ($\lambda = .967$, $F(8, 244) = 1.53$, $p > .05$, and $\lambda = .986$, $F(8, 244) = .442$, $p > .05$). As a result, gender was removed from the first model, and the refined model was tested again. The refined first MANOVA model yielded a significant difference between user role model and user shape avatars on the dependent variables ($\lambda = .778$, $F(8, 248) = 8.841$, $p < .001$). Also, the tests of between subjects effects detected that avatar types are significantly different on avatar rating ($F(1, 255) = 59.97$, $\eta^2 = .19$, $p < .001$). Figure 5-7 shows that the participants who were in the user role model condition had higher avatar ratings compared to the participants who were in the user shape condition, and this difference is statistically significant.

Avatar and Race

There was a significant interaction between avatar type and race. African American participants had higher game affect, and marginally higher progress ratings in the role model condition. White participants had lower engagement in the role model condition. South Asian participants had higher engagement in the role model condition. In our second MANOVA model where avatar type and race were the independent variables with a set of eight dependent variables, the test result indicated an interaction effect

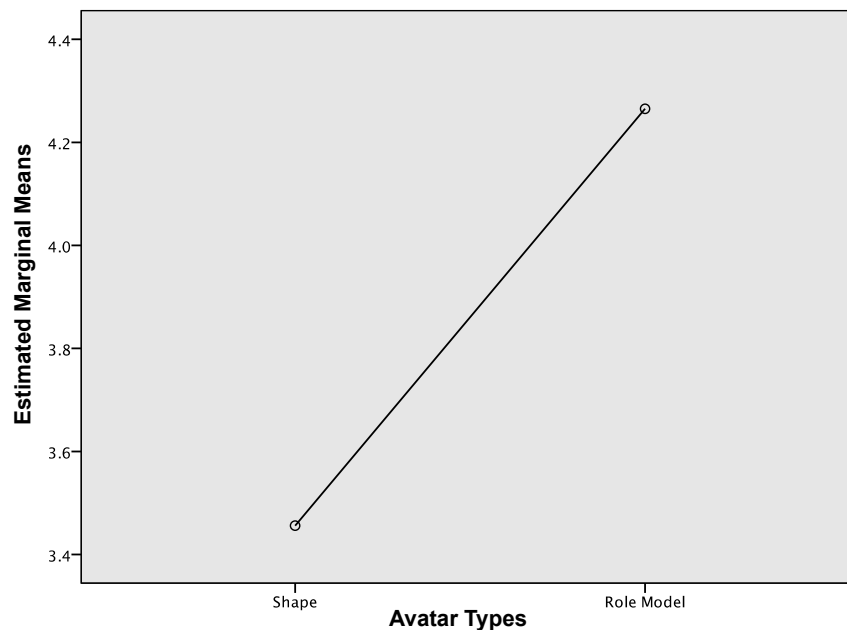


Figure 5-7: Avatar Ratings.

of avatar and race on the set of dependent variables (Pillai's Trace = .165, $F(24, 726) = 1.76$, $p < .05$). Tests of between subject effects showed that the interaction effect of avatar type and race has a significant effect on the average enjoyment rating ($F(3, 247) = 4.05$, $\eta^2 = .05$, $p < .05$) and progress rating ($F(3, 247) = 3.40$, $\eta^2 = .04$, $p < .05$). Independent-samples t-tests revealed that African American participants had higher game ratings (mean difference = 1.13, $p < .05$), and marginally higher progress ratings (mean difference = 0.68, $p < .1$) in the role model condition. White participants reported *lower* engagement in the role model condition (mean difference = 0.45, $p < .01$). South Asian players reported higher engagement in the role model condition (mean difference = 0.55, $p < .05$). See Figure 5-8 for illustration.

Performance Split

Players that completed few levels collected more bonus items in the role model condition. Players that completed ≤ 1 levels collected more bonus items in the role model condition, $p < .05$. See Table 5.8.

Same vs. Different Gender Role Models

No significant differences were found. See Table 5.9.

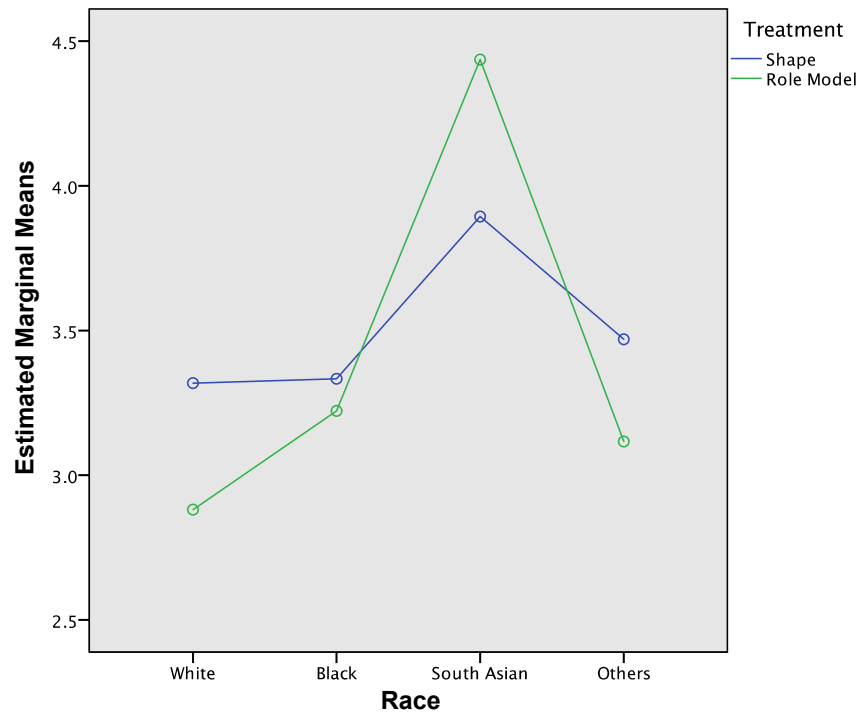


Figure 5-8: Average Enjoyment.

Table 5.8: Players completing ≤ 1 levels.

Attribute	<i>R</i> -Mean	<i>R</i> -SD	<i>S</i> -Mean	<i>S</i> -SD	t-test
Levels Completed	0.52	0.50	0.44	0.50	1.05
Total Bonus Items	0.83	1.33	0.39	1.00	2.56*
Average Enjoyment	3.23	1.34	3.36	1.11	0.49
Avatar Rating	4.11	0.86	3.55	0.95	4.28**

* $<.05$, ** $<.01$, R = Role Model, S = Shape, SD = Standard Deviation

Table 5.9: Participants selecting same gender role models versus different gender role models.

Attribute	<i>Rs</i> -Mean	<i>Rs</i> -SD	<i>Rd</i> -Mean	<i>Rd</i> -SD	t-test
Levels Completed	1.43	1.07	1.23	1.11	1.00
Total Bonus Items	2.23	2.90	1.67	2.55	1.10
Average Enjoyment	3.43	1.20	3.32	1.18	0.43
Avatar Rating	4.21	0.84	4.08	0.77	0.91

* $<.05$, ** $<.01$, *Rs* = Same Gen., *Rd* = Diff. Gen., SD = Standard Deviation

Table 5.10: Participants selecting same race role models versus different race role models.

Attribute	<i>Rs</i> -Mean	<i>Rs</i> -SD	<i>Rd</i> -Mean	<i>Rd</i> -SD	t-test
Levels Completed	1.38	1.11	1.39	1.03	0.03
Total Bonus Items	2.27	3.06	1.84	2.39	1.00
Average Enjoyment	3.47	1.21	3.32	1.17	0.70
Avatar Rating	4.31	0.79	3.97	0.83	2.73**

* $<.05$, ** $<.01$, *Rs* = Same Race, *Rd* = Diff. Race, SD = Standard Deviation

Gender Selections:

A chi-square test was used to determine whether there was a significant difference between male and female participants and the chosen role model's gender. 76.5% of female role models were chosen by female participants. 79.1% of male role models were chosen by male participants. The difference between male and female participants was significant, $\chi^2 = 39.63$, $df = 1$, $p < .001$.

Same vs. Different Race Role Models

Players had higher avatar ratings for same race role models. Players reported a higher avatar rating for same race role models, $p < .01$. See Table 5.10.

Race Selections:

A chi-square test was used to determine whether there was a significant difference in the chosen role model's race. Participants tended to pick a similar race role model. The difference was statistically significant, $\chi^2 = 115.52$, $df = 9$, $p < .001$.

Between Role Model Types

A cross tabulation was checked for any difference between participants' performance across role model types. The Chi-square test indicated no significant difference ($\chi^2 = 4.29$, $df = 11$, $p > .05$) between high and low performing participants across the 12 different role model types (see Table 5.11).

Table 5.11: Performance by role model types

Role Model Type:	High/Low performing groups		Total
	Low Performing	High Performing	
Actor	61.1%	38.9%	100.0%
Scientist	37.5%	62.5%	100.0%
TV personality	50.0%	50.0%	100.0%
Astronaut/Pilot	50.0%	50.0%	100.0%
Athlete	48.1%	51.9%	100.0%
Author	50.0%	50.0%	100.0%
Fictional Character	56.4%	43.6%	100.0%
Magnate	40.0%	60.0%	100.0%
Musician/Singer	61.9%	38.1%	100.0%
Political Figure	58.8%	41.2%	100.0%
Religious Figure	75.0%	25.0%	100.0%
Other	44.4%	55.6%	100.0%

Experiment-Specific Discussion (Shape vs. RoleModel)

The results suggest that role model avatars can enhance performance and engagement for some groups of participants. African American participants had higher game affect in the role model condition. South Asian participants had higher engagement in the role model condition. Participants completing ≤ 1 levels had higher performance in the role model condition. Therefore, an AI system that generates avatars would do well to utilize both the player demographics and the avatar type.

Participants in the user role model condition rated same race role models as higher. This, and the general trend observed in Tables 5.9 and 5.10, supports the literature, i.e., role models of similar gender and race appear to be more effective [52, 356]. Unexpectedly, we found that white participants had *lower* reported engagement in the role model condition. Previous studies have reported that role models can improve the academic performance of some social groups (i.e., female and African American participants), so while we would expect little to no effect in white participants, we see the *opposite* effect. While our current analyses cannot conclude as to *why* this occurred, this is an indication that role models

may not *always* be effective. For instance, one possible alternative explanation is cultural differences (e.g., a human photo may appear out of place). In addition, “superstar” role models can cause self-deflation [341].

Indeed, this is a complex topic; if we had attempted to link effects to specific types of role models (e.g., scientists, athletes, etc.), it is unclear whether the effect is due to the *type of role model*, the *type of person that picks that type of role model*, or *both*. More targeted studies are needed to explore specific role models. Our results expand upon the findings of the social science literature to also include role model *avatar* as a possible means of enhancing player engagement and performance.

5.1.9 Shape vs. Scientist vs. Athlete

Previous Experiment: [Shape vs. RoleModel](#)

Category: [Central Avatar Experiments](#)

Next Experiment: [Successful Likeness](#)

Experiment Overview (Shape vs. Scientist vs. Athlete)

Here, we describe an experiment (N = 890) exploring the use of (a) scientist role models, (b) athlete role models and (c) simple geometric shapes, as game avatars. Using the Game Experience Questionnaire (GEQ) [249], we find that over all participants, the use of avatars that looked like scientist and athlete role models led to highest flow and immersion. For female participants, the use of scientist avatars led to highest immersion and positive affect, and lowest tension and negative affect. The results here indicate that role model avatars have the potential to positively affect player game experience. This may especially be impactful for educational games, in which higher engagement could in turn influence learning outcomes [56].

Experiment-Specific Methods (Shape vs. Scientist vs. Athlete)

Our experiment aims to compare three avatar types: (a) scientist role models, (b) athlete role models, and (c) simple geometric shapes. The goal is to see if participants of different avatar type have differing game performance and game experience as measured by the GEQ. We strongly suspected ahead of time that (1) scientist avatars would outperform athlete and shape avatars, and (2) athlete avatars would outperform shape avatars. The experiment takes place in *Mazzy* ([278]; Section 3.1).

Avatar Conditions

The three avatar conditions we tested were:

- a. Scientist Avatars
- b. Athlete Avatars
- c. Shape Avatars

In each condition, players selected (inside the game) from a pool of eight possible choices. The pool of role models is composed of famous individuals, selected for diversity (i.e., exactly half the role models are female, and exactly half the role models are black or African American). When a user selects an avatar, there is a three-sentence summary presented of the avatar (e.g., “You’ve selected Albert Einstein. Albert Einstein was a German-born theoretical physicist., etc.). These were taken verbatim from their Wikipedia article. Avatars are always presented in a randomized ordering on the screen (see Figures 5-9, 5-10 and 5-11). Inside the game, the avatar consists of a 60 x 60 pixel game character that moves according to the user’s programs.

Quantitative and Qualitative Measures

For performance, we only analyze the number of levels completed by players. For measuring game experience, we use the GEQ [249]. We also included a single, 5-item Likert scale question on how the user felt towards the game character (1:Strongly Negative to 5:Strongly Positive).

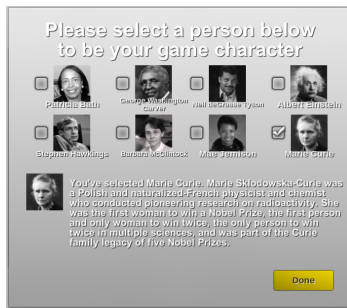


Figure 5-9: Scientists.



Figure 5-10: Athletes.

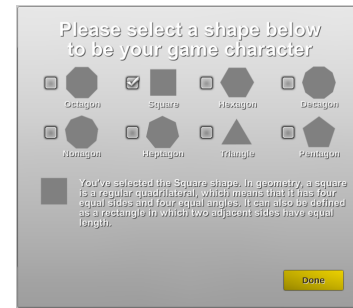


Figure 5-11: Shapes.

Participants

890 participants were recruited through Mechanical Turk. The data set consisted of 528 male, and 362 female participants. There were 712 white participants, 61 black or African American, 29 Chinese, 15 Filipino, 9 Korean, 8 American Indian, 8 Asian Indian, 8 Vietnamese, 8 Other Asian, 3 Japanese, 1 Native Hawaiian, 1 Guamanian or Chamorro, and 27 other. Participants were between the ages of 18 and 75 ($M = 31.4$, $SD = 9.0$), and were all from the United States. Participants were reimbursed \$1.50 to participate in this experiment.

Design

A between-subjects design was used: avatar type was the between-subject factor. Participants were randomly assigned to a condition.

Protocol

Prior to starting the game, players were informed that they could exit the game *at any time* via a red button in the corner of the screen. When participants were done playing (either by exiting early, or by finishing all 12 levels), participants returned to the experiment instructions, which then prompted them with the GEQ and then a demographics survey.

Results & Findings (Shape vs. Scientist vs. Athlete)

We find that using scientist avatars resulted in the highest scores on *immersion*, and that scientist and athlete avatars resulted in the highest scores on *flow*. For female

Table 5.12: Overall level completion statistics.

Avatar Condition	N	Mean	SD
Scientist	278	7.47	3.02
Athlete	308	7.40	2.87
Shape	287	7.16	2.84

Table 5.13: Female participant level completion statistics.

Avatar Condition	N	Mean	SD
Scientist	108	7.57	3.15
Athlete	122	7.00	3.00
Shape	128	7.02	2.93

participants, scientist avatars resulted in the highest scores on *immersion*, and *positive affect*, and in the lowest scores on *tension* and *negative affect*. We observe that there is consistent ordering in that scientist avatars are better than athlete avatars, and athlete avatars are better than shape avatars, across all subjective measures.

Aggregate

17 of the participants completed 0 levels. 3 of these were in the athlete condition, 4 in the scientist condition, and 10 in the shape condition. Being that the very first level of the game requires only following a set of simple directions (a basic tutorial level), these participants invested minimal effort and provide data of limited use. These participants are therefore excluded from further analysis. We report on data from 873 participants. The one-way ANOVA found no significant effect of avatar type on levels completed, $F(2, 870) = 0.89$, $p = 0.41$ (see Table 5.12).

A MANOVA revealed a statistically significant difference in GEQ responses and avatar ratings based on the participant's avatar type, $F(86, 1654) = 2.23$, $p < .0001$; Wilk's $\lambda =$

0.803, partial $\eta^2 = .10$. See Figure 5-12. Univariate testing found the effect to be significant for the following items: “I was deeply concentrating on the game” (*flow*), $F(2, 870) = 5.73$, $p < 0.005$; “I was fully occupied with the game” (*flow*), $F(2, 870) = 3.36$, $p < 0.05$; “It was aesthetically pleasing” (*immersion*), $F(2, 870) = 5.63$, $p < 0.005$; and “I felt imaginative” (*immersion*), $F(2, 870) = 4.30$, $p < 0.05$. Avatar rating was also found to differ between conditions, $F(2, 870) = 5204$, $p < 0.0001$.

In order to compare the effects of avatar type on these measures, we additionally calculated posthoc comparisons (LSD) between all conditions. The pair-wise comparisons revealed that the scientist condition GEQ rating was higher on “It was aesthetically pleasing” (*immersion*), $p < 0.05$, and “I felt imaginative” (*immersion*), $p < 0.005$, than the athlete condition. The scientist condition was also higher on “I was fully occupied with the game” (*flow*), $p < 0.05$, and “It was aesthetically pleasing” (*immersion*), $p < 0.005$, than the shape condition. The athlete condition GEQ rating was higher on “I was deeply concentrating on the game” (*flow*), $p < 0.0001$, and “I was fully occupied with the game” (*flow*), $p < 0.05$, than the shape condition. The scientist avatar rating was higher, $p < 0.0001$, than the athlete and shape condition. The athlete avatar rating was higher, $p < 0.0001$, than the shape condition.

Female Participants

The one-way ANOVA found no significant effect of avatar type on levels completed by female participants, $F(2, 355) = 1.30$, $p = 0.27$ (see Table 5.13).

The MANOVA revealed a statistically significant difference in GEQ responses and avatar ratings based on the participant’s avatar type, $F(86, 626) = 1.31$, $p < 0.05$; Wilk’s $\lambda = 0.718$, partial $\eta^2 = .15$. See Figure 5-13. Univariate testing found the effect to be significant for the following items: “I was interested in the game’s story” (*immersion*), $F(2, 355) = 3.87$, $p < 0.05$; “It was aesthetically pleasing” (*immersion*), $F(2, 355) = 3.41$, $p < 0.05$; “I felt imaginative” (*immersion*), $F(2, 355) = 3.38$, $p < 0.05$; “I felt irritable” (*tension*), $F(2, 355) = 4.95$, $p < 0.01$; “I could laugh about it” (*positive affect*), $F(2, 355) = 3.28$, $p < 0.05$; and “I was distracted” (*negative affect*), $F(2, 355) = 3.34$, $p < 0.05$. Avatar rating was also found to differ between conditions for female participants, $F(2, 355) = 19.49$, $p < 0.0001$.

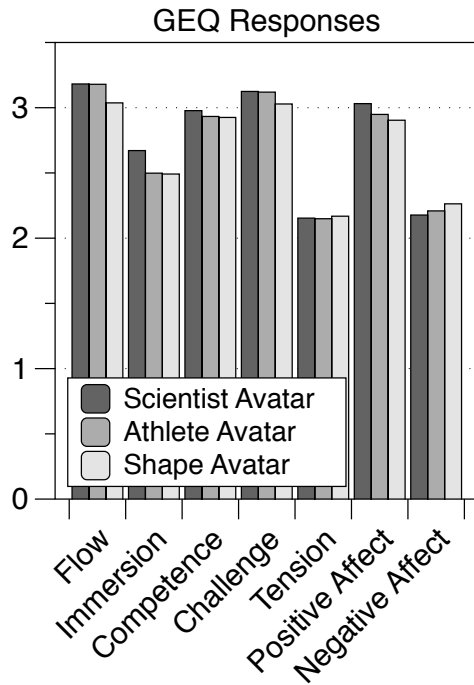


Figure 5-12: Game Experience Questionnaire (GEQ) responses for all participants.

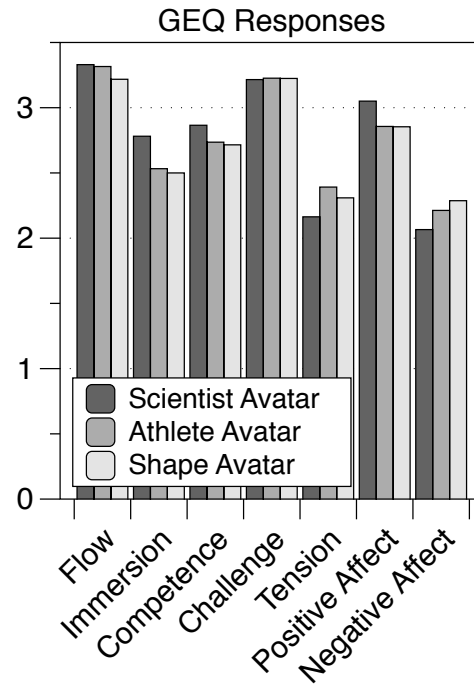


Figure 5-13: Game Experience Questionnaire (GEQ) responses for female participants.

In order to compare the effects of avatar type on these measures, we again additionally calculated posthoc comparisons (LSD) between all conditions. The pair-wise comparisons revealed that the scientist condition was higher on “I felt imaginative” (*immersion*), $p < 0.05$, and “I could laugh about it” (*positive affect*), $p < 0.05$, and lower on “I felt irritable” (*tension*), $p < 0.005$, and on “I was distracted” (*negative affect*), $p < 0.05$ than the athlete condition. The scientist condition was higher on “I was interested in the game’s story” (*immersion*), $p < 0.01$, “It was aesthetically pleasing” (*immersion*), $p < 0.01$, and “I could laugh about it” (*positive affect*), $p < 0.05$, and lower on “I felt irritable” (*tension*), $p < 0.05$, and “I was distracted” (*negative affect*), $p < 0.05$, than the shape condition. There were no significant differences between the athlete and shape conditions. The scientist avatar rating was higher, $p < 0.0005$, than the athlete and shape conditions. The athlete avatar rating was higher, $p < 0.05$, than the shape condition.

Experiment-Specific Discussion (Shape vs. Scientist vs. Athlete)

In the results, we have seen that using scientist avatars resulted in the highest scores on *immersion*, and that scientist and athlete avatars resulted in the highest scores on *flow*. For female participants, scientist avatars resulted in the highest scores on *immersion*, and *positive affect*, and in the lowest scores on *tension* and *negative affect*. We observe that there is consistent ordering in that scientist avatars are better than athlete avatars, and athlete avatars are better than shape avatars, across all subjective measures.

These results corroborate the findings in the social sciences, and demonstrate that those same findings are likely applicable via avatars inside educational games. For example, there have previously been three criteria for determining the effectiveness of a role model for boosting academic performance. (1) First, the role model should be perceived as competent [358]. Clearly, both the scientist and athlete avatars fulfill this condition. (2) Second, the role model should be perceived as an ingroup member, for instance the same gender or race [340, 356, 372]. This is also made possible for female and African American participants by having selected for diverse role models. (3) Third, the role model's record of success in domains where the role model's group is negatively stereotyped should be readily available [78, 357, 359]. To facilitate this, we have used only role models that are well known, and also provide a paragraph describing their achievements. Given that the experimental setting is a STEM education game, it is not of surprise that we have found scientist role models to be the more effective avatar type (e.g., [357]).

5.1.10 Successful Likeness

Previous Experiment: [Shape vs. Scientist vs. Athlete](#)

Category: [Central Avatar Experiments](#)

Next Experiment: [Phantoms vs. Non-Phantoms](#)

Experiment Overview (Successful Likeness)

Avatar research has almost exclusively explored avatars that remain the same regardless of context. However, there may be advantages to avatars that change during use. A plethora of work has shown that avatars personalized in one's likeness increases identification, while object-like avatars increase detachment. We posit that in certain situations within a game it may be more advantageous to have increased identification, while in other situations increased detachment. We present a study on *dynamic avatars*, or avatars that change types based on game context. In particular, we investigate what we term the *successful likeness* avatar. The *successful likeness* is an avatar that is only a likeness when the player is in a win state and at all other times an object. Our goal is to determine if this type of avatar can foster an increase in user performance and engagement. Our experiment (N=997) compares four avatars: 1) Shape, 2) Likeness, 3) Likeness to Shape, and 4) Shape to Likeness (*successful likeness*). We found that players using a *successful likeness* avatar had significantly better performance (levels completed) than all other conditions. Players using a *successful likeness* avatar had significantly higher play time (minutes played) than all other conditions. This was one of the most compelling results from our experiments. Additionally, we propose a theoretical model in which identification facilitates vicarious outcomes and in which detachment facilitates outcome dissociation. As performance and engagement are correlated to learning [225], *successful likeness* avatars may be crucial in educational games.

Experiment-Specific Background (Successful Likeness)

The *persona effect* was one of the earliest studies that revealed that the mere presence of a life-like character in a learning environment increased positive attitudes [323]. A wealth of empirical research since then has demonstrated that virtual characters are more influential when they have similar competencies [299], a similar gender [32, 202], and a similar ethnicity/race [438, 464]. However, research on the visual form and look of animated agents is sparse; it has been proposed that the following five dimensions are understudied: 1) the degree of "humanoidness," 2) the degree of stableness versus changeability of appearance

(morphing), 3) the degree of animation, 4) the degree of 3-dimensionality, and 5) the degree of realism [204]. The reason for the sparsity of research in this area has been attributed to two possible causes: 1) these questions are difficult to answer using existing methodologies, and 2) people do not readily accept the idea that appearance affects our intellectual processes (e.g., “Don’t judge a book by its cover”) [204]. In this work, we propose to explore the *changeability of appearance* aspect of avatars.

In particular, our work is based on increasing evidence that demonstrates that abstract avatars increase player-avatar detachment via low identification (my avatar is *not* me), high sense of control (my avatar is like a tool), low sense of responsibility (my avatar has no needs), etc. [25, 63, 282]. Work in human-computer interaction, psychology, and marketing suggests that within virtual environments, success and failure is attributed to avatars and through them also affects users [83, 243, 386, 546, 558]. These effects are more powerful with avatars with whom we identify [141, 528]. Here, we perform the first study to our knowledge on *dynamic avatars*. More specifically, we study what we call the *successful likeness* avatar, an avatar that is normally abstract (e.g., a shape), but that becomes a likeness in win states. Our goal is determining if selectively increasing and decreasing user identification with the avatar during key moments of the game experience can result in increased performance and engagement. We found that participants did not significantly differ in reported engagement between conditions. However, participants using a *successful likeness* avatar both completed significantly more levels, and played the game for a significantly longer period of time, suggesting greater performance and engagement (see [29] for predicting engagement via play time). Since both performance and engagement have been correlated to better learning outcomes in educational games [56, 225], our work has important implications for avatar design in educational environments.

Our work here is based on research on avatars, agents, and “blended identities” [218]. Although in this work we are studying avatars, an abundance of work on agents (i.e., virtual pedagogical agents, teaching agents, etc.) helps to guide our study. In particular, a large body of work has shown that avatars and agents that share users’ external characteristics (e.g., age, gender, race, clothing, etc.) are more influential and are linked to better learning outcomes

[18, 22, 32, 202, 265, 299, 438, 464]. This is posited to be a result of similarity-attraction, the theory that people are attracted to similar others [79, 256]. Functional neuroimaging has found that perceived similarity is an important factor in a person's ability to simulate the internal state of another person [378]. Likewise, Mobbs et. al found that when a participant watched a game show contestant with high perceived similarity, the participant experienced significant increases in both subjective and neural responses to vicarious reward [380]. Other work suggests that what is experienced by an avatar is also experienced by its user [83, 243, 386, 546, 558]. This effect is more powerful via avatars that we identify with [141, 528], identification being positively correlated to such factors as representation of emotions and intent [212], physical resemblance [346], and avatar customization [520].

In the past decade, it has become apparent that avatars play an important role in affecting our behaviors. The Proteus effect describes an individual's tendency to conform to behavior typically associated with how an avatar appears [561]. For example, two of the earliest studies found that participants with taller avatars were more aggressive, and that participants with more attractive avatars were more confident. Since avatars affect us in a subtle way, they are a form of "embedded content" [293], which studies have shown is more effective than "message-driven" agendas [66]. Avatars, or "blended identities," [218], can be pivotal in enabling our capacities to put ourselves inside other identities. However, the unfortunate consequence is avatars can also be used to reinforce stereotypes and perpetuate hegemonic views, e.g., women as victims of violence. Fortunately, some representations can begin to combat these stereotypes, e.g., playing a computer science learning game as Marie Curie [285]. For instance, research has shown that abstract (or object-like) representations, such as a geometric shape, lead to detachment with the avatar and outcomes associated to the avatar [25, 63, 224, 280]. Because of the potential usefulness in exploring the dichotomy between identification and detachment, we investigate the *successful likeness*. This avatar is abstract (shape) when the player is not in a win state (to facilitate detachment), and a likeness (Mii) when the player is in a win state (to facilitate identification). Our goal is to test if this type of avatar can enhance player performance and engagement.

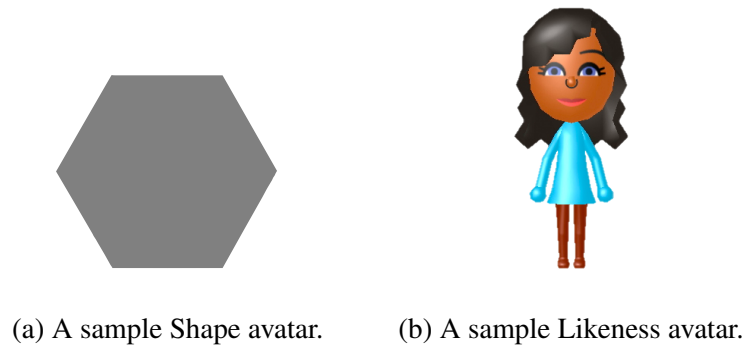


Figure 5-14: Sample avatars.

Experiment-Specific Methods (Successful Likeness)

Our experiment consisted of a between-subjects design. Our goal was to measure performance and engagement across conditions. The experiment takes place in *Mazzy* ([278]; Section 3.1).

Conditions: Our four avatar conditions were:

- 1) Shape
- 2) Likeness
- 3) Likeness to Shape
- 4) Shape to Likeness

Participants were all told that they would be playing a game. No other details were specified. Players were asked to use a publicly available customization system to create a Mii (the *Likeness*). A Mii is a type of avatar developed by Nintendo, chosen since Miis were designed with the intention that most users would create likeness avatars (the word “Mii” is a blend of “Wii” and “me”). Furthermore, players were told to create an avatar that looked like themselves. Players then picked out of eight possible geometric shapes (the *Shape*). Every player created a *Likeness* avatar and selected a *Shape* avatar (see Figure 5-14). If a participant was assigned to Condition 1, their avatar was always a shape. In Condition 2, their avatar was always a Mii. In Condition 3, their avatar was normally a Mii, but when a level was

successfully won, the avatar became a shape. In Condition 4, their avatar was normally a shape, but when a level was successfully won, the avatar became a Mii (*successful likeness*). The ‘winning’ avatar (a shape in Conditions 1 & 3, and a Mii in Conditions 2 & 4) was displayed centered in the middle of the screen. All other aspects of the experiment were identical across conditions.

Measures: Our performance measures consist of levels completed and time played, while our engagement measure is the Game Experience Questionnaire (GEQ) [249].

Participants: 997 participants were recruited through Mechanical Turk. The data set consisted of 560 male, and 437 female participants. Participants self-identified their races/ethnicities as white (665), Asian Indian (163), black or African American (55), American Indian (14), Chinese (13), Filipino (13), Korean (10), Japanese (6), Vietnamese (4) and other (54). Participants were between the ages of 18-72 ($M = 30.1$, $SD = 8.2$). Participants were reimbursed \$1.50 to participate in this experiment.

Design: Our design was a between-subjects design: avatar condition was the between-subjects factor. Participants were randomly assigned to a condition.

Protocol: Prior to starting the game, players were informed that they could exit the game at any time via a red button in the corner of the screen. When participants were done playing (either by exiting early, or by finishing all 12 levels), participants returned to the experiment instructions, which prompted them with demographics.

Analysis: Data was analyzed in SPSS using analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA). We ran one ANOVA using *levels completed* as the dependent variable, and one ANOVA using *time played* as the dependent variable. Our MANOVA used *GEQ items* as the dependent variable. In all cases, our independent variable is *avatar condition*. To be aligned with our research question, we asked participants after the experiment to rate how similar they felt their Mii was to themselves (1: *Very Dissimilar* to 5: *Very Similar*). We removed participants that reported a similarity less than 3 (189). Additionally, we removed 35 outliers according to the criteria in Hoaglin (1987). These 224

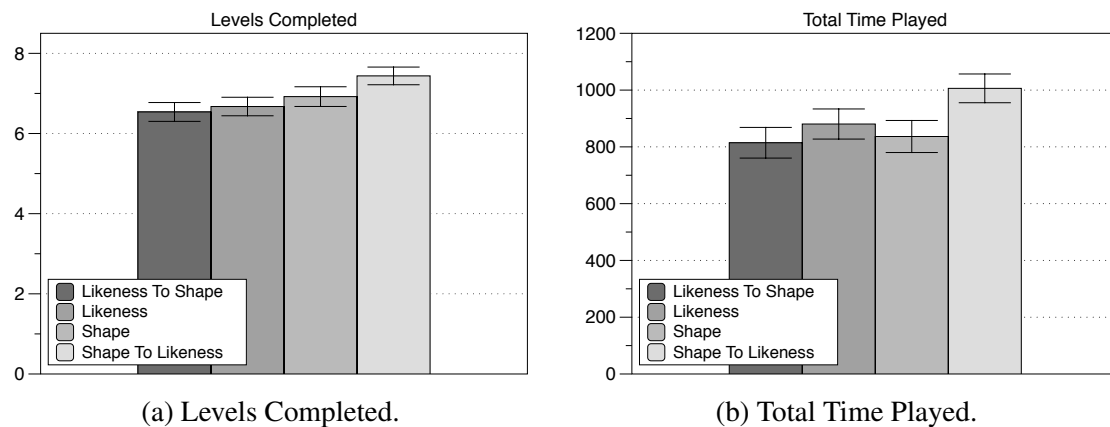


Figure 5-15: Performance.

participants were excluded from further analysis. Prior to running our MANOVA model, we checked the assumption of homogeneity of variance by Levene's Test of Equality of Error Variances, and the assumption was met by the data ($p > .05$). All reported p -values are two-tailed.

Results & Findings (Successful Likeness)

- Shape to Likeness participants were **highest performing**.
- Shape to Likeness participants spent the **most time in game**.
- **No significant differences** in GEQ responses.

An ANOVA revealed that *levels completed* was significantly different across avatar conditions, $F(3, 769) = 3.02, p < .05$. Post-hoc comparisons (LSD) revealed that the condition *Shape to Likeness* significantly outperformed *Likeness*, $p = .017$. The condition *Shape to Likeness* also significantly outperformed *Likeness to Shape*, $p = .007$ (see Figure 5-15a).

Similarly, an ANOVA revealed that *time played* (seconds) was significantly different across avatar conditions, $F(3, 769) = 2.69, p < .05$. Post-hoc comparisons (LSD) revealed that the condition *Shape to Likeness* had significantly longer play time than *Shape*, $p = .019$. The condition *Shape to Likeness* also had significantly longer play time than *Likeness to Shape*, $p = .010$. The condition *Shape to Likeness* had marginally longer play time than *Likeness*, $p = .072$ (see Figure 5-15b).

A MANOVA revealed that there was no statistically significant difference in GEQ responses across avatar conditions, $F(126, 2190) = 1.02$, $p = .43$; Pillai's Trace = .17, partial $\eta^2 = .055$. See Figure 5-16.

None of the participants correctly guessed the purpose of the experiment.

Experiment-Specific Discussion (Successful Likeness)

We found that the *Shape to Likeness (successful likeness)* condition had significantly increased **Levels Completed** and **Time Played**. GEQ responses did not significantly differ. These results support our initial hypothesis that having a shape avatar (greater detachment) when the player is not in a win state, and having a likeness avatar (increased identification) when the player is in a win state, would outperform other avatar types. The worst performing condition was the inverse condition: *Likeness to Shape*.

What do these results mean? For example, the *successful likeness* condition participants on average completed about 1 more level and played for about 3.2 minutes longer than *Likeness to Shape* condition participants. Longer game playtime can be used as a measure of engagement [29]. Moreover, both increased game performance and engagement have been correlated to better learning outcomes in educational games [56, 225]. Therefore, *these results are suggestive that, over longer periods of time, dynamic avatars could have beneficial effects on players.*

Why did this happen? Multiple disciplines have independently demonstrated that avatars with a higher degree of perceived similarity may better facilitate vicarious experiences, positive or negative [18, 22, 32, 79, 83, 141, 202, 243, 256, 265, 299, 386, 438, 464, 528, 546, 558]. Moreover, neural imaging has demonstrated that watching a similar person experience a reward also increases our own vicarious reward [380]. Effects of an avatar similar to oneself may persist even after the experiment. Fox & Bailenson found that watching one's avatar exercising resulted in significantly more exercise on the part of the participant, 24 hours later, as compared to participants that watched one's avatar loitering or

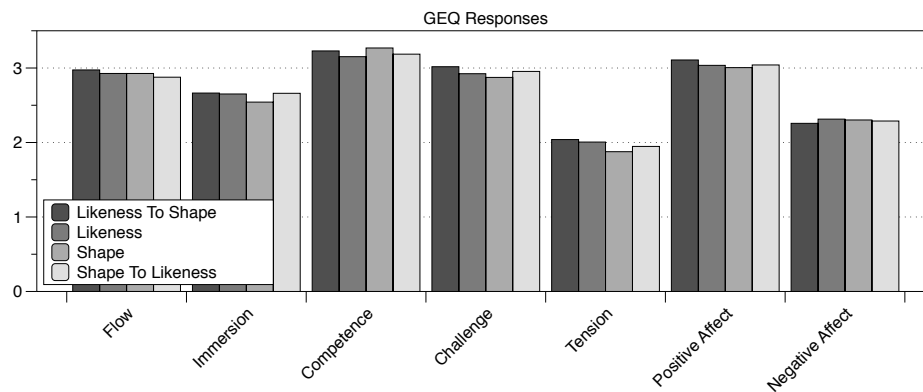


Figure 5-16: Game Experience Questionnaire (GEQ) responses for all participants.

a virtual other exercising [167]. Lastly, abstract (or object-like) avatars can better facilitate detachment and may play a role in helping users dissociate from failure outcomes, such as in cases requiring “debugging” [25, 63, 224, 280].

How generalizable are these results? The work here was a single study of how dynamic avatars affected engagement and performance for 997 participants in a coding game. While we feel the results are well supported by the literature, there should be additional investigation of the specific physiological effects of dynamic avatars. While one possible approach is to ask participants questions from, e.g., the Player-Avatar Interaction (PAX) questionnaire [26], we feel that post-game surveys will be a difficult approach given the rather subtle differences in the experience between conditions. While these subtle differences manifested as tangible differences in performance, they did not manifest in tangible differences in reported engagement. Even if players differed on some self-report (e.g., “This avatar understands me.”), it’s not readily apparent how we can disambiguate, in the dynamic avatar case, between the non-win state avatar, the win state avatar, or some combination. Because these avatars are different than any avatar previously studied, we will need new methods to study them. We find some parallels in work on multiple agents; for instance, it has been found that multiple virtual pedagogical agents with ‘compartmentalized’ roles (e.g., one agent provides confidence-boosting messages, another provides information support, etc.) provide significantly better learning outcomes than a single agent [33, 34, 403]. Here, we instead have multiple avatars, and we are facilitating either greater identification or greater

detachment depending on the game context. We plan to further investigate this phenomenon in the near future. We are partnered with a non-profit and will be studying Computer Science learning using these avatars in Cambridge schools. We aim to use EEG devices, e.g., the EPOC+, to measure brain activity of participants over the course of game play. Ultimately, such an approach would help us determine the specific physiological influences of these *dynamic avatars*.

Increasingly, there has been research on the different external characteristics of avatars and agents and how they affect users in educational environments [32, 202, 299, 323, 438, 464]. However, their visual form and look has been understudied, including avatars that change from one form to another (morphing) [204]. Here, we provide the first study to our knowledge on *dynamic avatars*, or avatars that are different depending on whether the user is in a win state or not. We found that the dynamic avatar, *successful likeness*, outperformed all other conditions in terms of levels completed. These same participants also played the game significantly longer. We posit that this is a result of shapes (abstract) as avatars leading to more detachment and Mii avatars (likeness) leading to more vicarious experience. Educational systems and games could benefit greatly from such a model of representation, shielding users from internalizing failure, and basking them in self-success-identification.

5.2 Additional Avatar Experiments

Experiment listing:

[Phantoms vs. Non-Phantoms](#)

[Red vs. Blue](#)

5.2.1 Phantoms vs. Non-Phantoms

Previous Experiment: [Successful Likeness](#)

Category: [Additional Avatar Experiments](#)

Next Experiment: [Red vs. Blue](#)

Experiment Overview (Phantoms vs. Non-Phantoms)

Avatar apparitions (or “ghosts”) are visual manifestation of other players’ avatars that are non-interactive such as used in popular videogames, e.g., *Dark Souls*. As such, these virtual identities are deployed unlike avatars in either single player or multiplayer. Avatar apparitions can be used to reveal patterns, mistakes, and successes of other players. In this manner they can act as effective tutorial supports. They are often displayed during real-time play, however they can be either live or based upon previous play experiences of other users. We performed a study (N=523) exploring the effects. Players were randomly assigned to a vanilla educational game, or the same game plus apparitions. Apparitions were found to increase performance, but decrease engagement. Apparition game players used significantly less hints, and reported lower challenge. Qualitative results suggest novice players have higher affect towards apparitions. These results can better inform design of educational games with unobtrusive tutors.

Experiment-Specific Background (Phantoms vs. Non-Phantoms)

Avatar¹ apparitions are an innovation in videogaming that non-interactively convey experiences of other players. Since these apparitions have been used as tutorial supports in popular videogames (particularly the notoriously challenging *Dark Souls* games [173], see Figure 5-17) we began to consider whether the approach might be useful in educational games. A benefit is that students who are resistant to overly didactic direct instruction might find avatar apparitions to be effective as more unobtrusive tutorial supports for performance (number of tries, time spent solving problems, time spent overall in the game, etc.) and engagement (self-reported positive affect toward and involvement in the game).

We decided to run an experiment implementing these apparitions in *Mazzy*. The apparitions

¹The terms avatar and player character will be used interchangeably, although we consider avatar to be the more general term as it is not limited to game/play settings.



Figure 5-17: In the videogame *Dark Souls* two apparitions (white figures on the left and right) show the player (center) what happens when the player character runs directly at the dragon.

consist of random playthroughs of other players for the current level (failures and successes). However, players only see the movement of other players, never the actual code that was used to create those movements. This is analogous to seeing another person’s program output (whether it was a failure or success), but not that person’s code. In our experiment (N=523), we find that apparitions increased player performance, but reduced game affect. Apparitions appear more beneficial for novice players. Because apparitions are inexpensive, simple to maintain, and can improve performance (which has been linked to learning [225]), this topic can inform future design of games.

While results will be discussed at more length below, as motivation consider the following response from our online study of 523 players to the question: “What did you think of [other characters’ apparitions]?”

Participant No. 432 said:

This may sound silly, but they reminded me of the brief visions of other players that one sees when playing the game *Dark Souls*; other characters who are playing in the world at the same time and place as you can be seen as brief

phantoms. This helps Dark Souls to feel more engaging and more "full of life"; it's a nice bit of companionship. I felt the same way about these other characters, because they reminded me of that.

While Gekker [185] describes Dark Souls:

The game is single player in its core, but [...] occasionally apparitions of other players appear out of thin air, doing battle with enemies unseen to your character. You cannot interact with them in any way, and their sole purpose is moral support of sorts- reminding you that there are others out there fighting the same war as you do. [...]

Dark Souls, *Dark Souls 2*, and *Bloodborne* are developed by FromSoftware; combined sales exceed 10M copies. Even though these games all use apparitions, this is not often extensively discussed in game studies or popular journalism about videogames and to our knowledge is largely unexplored in educational gaming.

That being said, the use of multiplayer game modes is well-studied in the literature. For instance, multiplayer game modes have been found to increase physical exertion when compared to single player game modes [404, 426]. Apparitions, however, fall somewhere in-between single-player and multi-player, since it is neither a completely solo experience, nor is there any interaction with other players. The term "asynchronous multiplayer" [57] refers to sequenced game play. A turn-based game such as *Diplomacy*, a strategy war game in which players play-by-mail by sending moves to a central game master, is a good example. Game replays are increasingly a topic of research interest, however they are usually distinct from gameplay itself. While apparitions share some properties of both asynchronous multiplayer and replays, neither term is adequate. Rather, apparitions are random instantiations of other players' movements inside a game, during the game itself. Apparitions can be thought of as having a clear relationship to intelligent tutors, educational systems, etc. in that they are a form of cognitive support or scaffold (e.g., [568]) for the player.

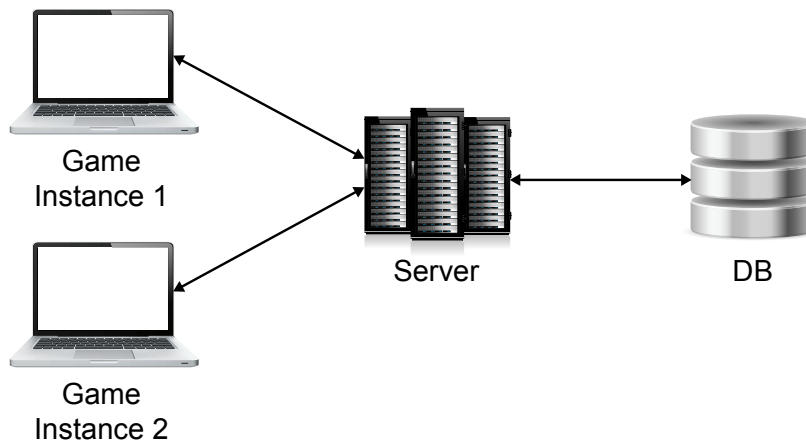


Figure 5-18: Apparitions network.

Experiment-Specific Methods (Phantoms vs. Non-Phantoms)

The experiment takes place in *Mazzy* ([278]; Section 3.1). In our game, apparitions works as follows (see Figure 5-18). When a user runs a program, a message is sent to the server to save that *run*. The *run* contains the level, the program, and all avatar data. The server saves the *run* in a database. During the game, user clients query the server at random intervals (e.g., `Random.Range(1f,35f)` seconds). The server sends the client recent *runs*. Only *runs* for the current level, and from others, are sent. Out of these *runs*, one is picked at random. When an apparition appears, it runs an identical program to the one that was run on its host computer. Apparitions appear exactly as they did in its host’s computer (all customization data such as colors, hair, etc.). In order to distinguish them, their alpha is set to 50%.

There are a few things to note. First, we have relaxed the real-time constraint for apparitions. This means apparitions are viable without simultaneous users. When users *are* concurrent, apparitions are real-time. Second, there are a few arbitrary details. These include the random interval spawn, apparition alpha, etc. These were deliberated on and chosen to balance a close simulation to other games, and our game’s context.²

Conditions

²Prior to the experiment, the system was first “seeded” by ten complete playthroughs.

Our experiment aims to compare two conditions: (a) No Apparitions, and (b) Apparitions. The goal is to see if participants in the two conditions have different game performance and affect towards the game. Players in the Apparitions condition can see apparitions. In all other respects the game is identical.

The performance measures we recorded were:

- **Levels completed:** The number of levels completed.
- **Total attempts:** The number of attempts in each level.
- **Total hints:** The number of hints consumed.

In addition, we used the Game Experience Questionnaire (GEQ) [249] to measure positive affect, negative affect, challenge, etc.

Participants

523 participants were recruited through Mechanical Turk. The data set consisted of 274 male, and 249 female participants. Participants were between the ages of 19 and 74 ($M=32.1$, $SD=9.9$), and were all from the United States. Participants were reimbursed \$1.50 to participate in this experiment.

Design

A between-subjects design was used: apparitions was the between-subject factor. Participants were randomly assigned to a condition.

Experiment Protocol

Prior to starting the game, players were informed that they could exit the game at any time via a red button in the corner of the screen. When participants were done playing (either by exiting early, or by finishing all 12 levels), participants returned to the experiment instructions, which then prompted them with the GEQ and, for players in the Apparitions condition, two qualitative questions: 1) “What/Who did you think were the other game characters”, and 2) “What did you think of them?”.

Analysis

Data was extracted and imported into Statistical Package for Social Science (SPSS) version 22 for data analysis using multivariate analysis of variance (MANOVA). The dependent variables are *levels completed*, *total attempts*, *hints used*, *GEQ items*; and the independent variable is *apparitions*. All the dependent variables are continuous variables. For the independent variable, the apparition status (i.e., 0 = no apparitions; 1 = apparitions) is a dichotomous variable. To detect the significant differences, we utilized MANOVA. We also ran targeted independent-samples t-tests on number of hints per level. This was to look for inter-level differences. Results are reported as significant when $p < 0.05$ (two-tailed). Before running MANOVA, all the variables included in the analyses were checked, and there were 18 outliers detected (Hoaglin, 1987). These 18 outliers were excluded from further analysis. Prior to running our MANOVA, we checked both assumption of homogeneity of variance and homogeneity of covariance by the test of Levene's Test of Equality of Error Variances and Box's Test of Equality of Covariance Matrices; there was a single violation for 1 item in *positive affect* (Levene's Test $p < .05$). All other variables met the assumptions ($p > .05$ for Levene's Test, and $p > .05$ for Box's Test). As the sample size is equal and large, MANOVA is robust here; as an added precaution, we use Pillai's Trace and not Wilk's Lambda, which is robust under violation (Olson, 1976).

Results (Phantoms vs. Non-Phantoms)

Apparitions players had lower affect, but performed significantly better.

Quantitative

Our MANOVA was statistically significant, $p < .05$. Between subjects testing found apparitions scored lower on the challenge question "I felt stimulated", $p < .05$. Apparitions players had a lower score on the positive affect question "I enjoyed it", $p < .05$. Apparitions players had a higher score on the negative affect question "I was distracted", $p < .05$.

For hints used, unpaired t-tests found that players in the apparitions condition used signifi-

cantly less hints in Level 7, $p < .05$. Players in the apparitions condition used marginally less hints in Level 8, $p < .1$.

What/Who did you think were the other game characters?

A small majority of participants felt that the apparitions were bots (52.4%), while somewhat less felt they were other players (41.5%). The rest were unsure or gave other less common answers (e.g., that the characters were themselves).

What did you think of them?

Most players felt that the apparitions were helpful (33%). Other players felt neutral towards the apparitions (22%), irritated by them (11%), interested by them (8%), felt they were skilled (7%), or similar to themselves (4%). Smaller numbers of people felt that they were unhelpful (3%), confusing (3%), unskilled (3%), odd (3%), or some other response (3%).

Players that viewed apparitions as helpful vs. irritating

We additionally compared the two clusters of players that viewed the apparitions as helpful and irritating. Participants that were irritated by apparitions in general performed better³ than participants that were allegedly helped by apparitions. Moreover, participants that were irritated reported having a less challenging experience. Engagement was *higher*⁴ in the group of participants that said apparitions were helpful. This suggested to us that apparitions may be irritating to high-performers but helpful to lower-performers.

Experiment-Specific Discussion (Phantoms vs. Non-Phantoms)

We first summarize our findings:

- Apparitions participants had **lowest affect**.
- Apparitions participants were **highest performing**.
- Apparitions appear **better for novices, worse for experts**.

³Higher levels completed, less hints used

⁴Higher flow, affect, etc.

Apparitions players experienced less challenge, and used less hints. While the argument can be made that apparitions, in a sense, give away the solution, one important distinction is that they serve only to trace the path and do not reveal code. While this may be helpful to remind users of potential paths, both failure cases and success cases, apparitions may act as supports in other ways as well.

Participant No. 174 said:

Each character gave me motivation to try to solve each problem as fast as I could. I didn't feel it was a competition, but I knew if other people were solving each level then I could do it also.

Therefore, it is possible that apparitions can act as a motivational tool of sorts, a reminder that the levels are solvable. Apparitions also, however, reduced positive affect towards the game.

It made me mad. I like to figure things out myself. I don't like to feel like I've been given an advantage. [...]

The players that said they were irritated by the apparitions reported less challenge and had higher performance than the players that said they were helped by the apparitions. This suggests that apparitions may be helpful for some, in particular more novice players. Participant No. 424 said that the apparitions made the game "less lonely", suggesting that players "click for that feature". Such a toggle could make apparitions amenable to both novices and experts alike.

Apparitions may be an effective mechanic. Apparitions participants had higher performance, reported lower challenge, and many cited apparitions as making the game less lonely. However, some participants felt irritated by the apparitions. Results suggest apparitions are beneficial to novices. The overhead of implementing apparitions is low: a basic database and some network messages. Apparitions are both companions and reminders that a solution exists. The prudent use of apparitions in educational games can add "life", even to what appears a solitary struggle.

5.2.2 Red vs. Blue

Previous Experiment: [Phantoms vs. Non-Phantoms](#)

Category: [Additional Avatar Experiments](#)

Next Experiment: [Feedback Positive vs. Negative vs. Neutral vs. Nothing](#)

Experiment Overview (Red vs. Blue)

The color red has been shown to hinder performance, motivation, and affect in a variety of contexts involving cognitively demanding tasks [146, 190, 244, 271, 314, 329, 374, 376, 493]. Teams wearing red have been shown to impair the performance of opposing teams [136, 235, 254, 424], present even in online gaming [251]. Although color is strongly contextual (e.g., red-failure association), its effects are posited to be sub-conscious [171] and operate powerfully even on nonhuman primates, e.g., Rhesus macaques (*Macaca mulatta*) take food significantly less often from an experimenter wearing red [296]. Here, we present one of the first studies on avatar color in a single-player game. We compared players using a red avatar to players using a blue avatar. Using the Game Experience Questionnaire (GEQ) [249], we find that players using a red avatar had a decrease in competence, immersion and flow. Our results are of consequence to how we design and choose colors in single-player contexts.

Experiment-Specific Background (Red vs. Blue)

Over 120 years of research on color and its effects on humans have led to Color-In-Context (CIC) theory [145]. CIC has six premises: (1) Color carries meaning, i.e., color is more than aesthetics, (2) Color influences psychological functioning, e.g., colors are evaluated to be hospitable or hostile [80, 144, 316, 565], (3) Color effects are outside of conscious awareness [313, 337, 423], (4) Color meaning is both learned and intrinsic, i.e., paired color associations such as pink is feminine; color vision as an adaptation [91, 246, 258, 383], (5) Color perception influences affect, cognition, and behavior, and vice versa [73, 213, 379],

(6) Color is context-specific, e.g., pink is frequently viewed as feminine on a baby's blanket, but not on Bazooka bubble gum [523].

This gives us a framework for understanding how color may affect us in digital spaces. For example, most students in primary school are primed to associate red and failure in an evaluative context [374, 382, 439, 471]⁵. Moreover, red has associations with blood, danger, fire and anger. Red has been posited to be a distractor signal. Since Hill's seminal paper on the Athens Olympic Games in 2004 in which it was found that red-wearing competitors won more bouts than blue-wearing competitors in four different sports [235], there has been a plethora of research on color, motivation, and achievement. Later work found this work to be consistent in a variety of sporting domains [136, 235, 254, 424], and even in an online FPS game [251].

However, one gap in the literature is color in single-player contexts⁶. To fill this need, we performed a study comparing players using a red avatar to players using a blue avatar, inside an educational game of our own creation. Although there is some question to whether, in the context of a sporting event, the color red is affecting the wearer, the opponents, or the referees, past work has consistently shown that red reduces mood, affect and performance in cognitive-oriented tasks [146, 190, 244, 271, 314, 329, 374, 376, 493]. For example, Lichtenfeld et. al showed that even just peripherally noticing red (e.g., hidden in a question, in the copyright notes at the end of a page, etc.) can have similar effects [329]. For this reason, we hypothesized that, if there were to be any effect on performance and game experience, that it would favor the blue avatar over the red avatar.

Experiment-Specific Methods (Red vs. Blue)

Our experiment aims to compare two colors of avatar: (a) blue avatar, and (b) red avatar. The goal is to see if participants using the two colors of avatar have differing game performance

⁵This association is not necessarily true across culture. For instance, an upward rise in China's stock market is represented in red [261, 567].

⁶Multiplayer studies on color exist [134, 251, 425], as do a few on colored environments [268, 400], but there are few studies on avatar color in single-player games.

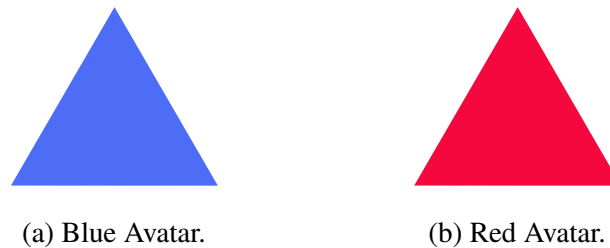


Figure 5-19: Avatars.

and game experience as measured by the GEQ. We strongly suspected ahead of time that results would favor the blue avatar. The experiment takes place in *Mazzy* ([278]; Section 3.1).

Avatar Conditions

The two avatar conditions we tested were:

- a. Blue Avatar
- b. Red Avatar

The avatar was a triangle shape in both conditions, colored either blue or red. Color is defined by lightness, chroma, and hue. We keep lightness and chroma constant using the Munsell color system [151]. Only colors that can be displayed with good accuracy on a computer screen were considered⁷. The specific colors chosen were 7.5PB 5/18 (▲) and 5R 5/18 (▲). See Figures 5-19a and 5-19b. Inside the game, the avatar consists of a 60 x 60 pixel game character that moves according to the user's programs.

Quantitative and Qualitative Measures

For performance, we only analyze the number of levels completed by players. For measuring game experience, we use the GEQ [249].

Participants

507 participants were recruited through Mechanical Turk. The data set consisted of 278

⁷<http://www.andrewwerth.com/aboutmunsell/>

male, and 229 female participants. Participants self-identified their races/ethnicities as white (407), black or African American (29), Asian Indian (24), Chinese (5), Korean (4), American Indian (3), Vietnamese (3), Japanese (2), Filipino (1) and other (29). Participants were between the ages of 18 and 65 ($M = 30.3$, $SD = 8.7$), and were all from the United States. Participants were reimbursed \$1.50 to participate in this experiment.

Design

A between-subjects design was used: avatar color was the between-subject factor. Participants were randomly assigned to a condition.

Protocol

Prior to starting the game, players were informed that they could exit the game *at any time* via a gray button in the corner of the screen. When participants were done playing (either by exiting early, or by finishing all 12 levels), participants returned to the experiment instructions, which then prompted them with the GEQ and then a demographics survey.

Analysis

Data was extracted and imported into Statistical Package for Social Science (SPSS) version 22 for data analysis using multivariate analysis of variance (MANOVA). The dependent variables were *GEQ items*; and the independent variable was *avatar color (blue or red)*. All the dependent variables are continuous variables. The independent variable avatar color (i.e., 0 = blue, 1 = red) was a dichotomous variable. To detect the significant differences between blue avatar and red avatar, we utilized one-way MANOVA. We also ran an independent-samples t-test on the variable *levels completed*. These results are reported as significant when $p < 0.05$ (two-tailed). Prior to running our MANOVA, we checked both assumption of homogeneity of variance and homogeneity of covariance by the test of Levene's Test of Equality of Error Variances and Box's Test of Equality of Covariance Matrices; and both assumptions were met by the data ($p > .05$ for Levene's Test, and $p > .001$ for Box's Test).

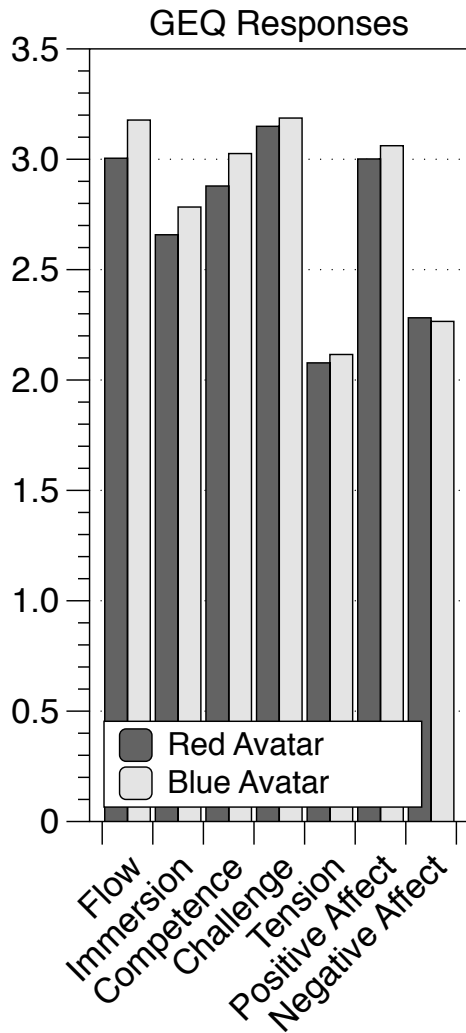


Figure 5-20: Game Experience Questionnaire (GEQ) responses for all participants.

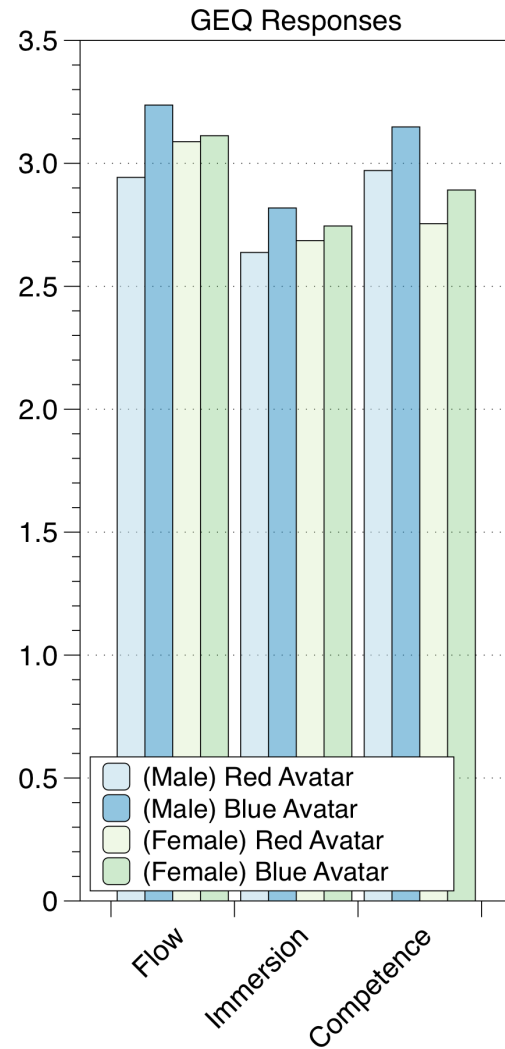


Figure 5-21: Game Experience Questionnaire (GEQ) responses for male and female participants.

Results & Findings (Red vs. Blue)

- Blue led to **higher flow** than Red
- Blue led to **higher immersion** than Red
- Blue led to **higher competence** than Red
- Blue led to **higher (avatar) affect** than Red

Aggregate

A MANOVA revealed a statistically significant difference in GEQ responses based on the

participant's avatar color, $F(42, 464) = 1.43$, $p < .05$; Wilk's $\lambda = 0.885$, partial $\eta^2 = .12$. See Figure 5-20. Pair-wise comparisons revealed that the blue avatar GEQ rating was higher on "I was fully occupied with the game" (*flow*), $p = .015$, "It felt like a rich experience" (*immersion*), $p = .018$, and "I felt competent" (*competence*), $p = .044$. Blue was marginally higher on "I felt completely absorbed" (*flow*), $p = .058$, "I forgot everything around me" (*flow*), $p = .077$, "I lost track of time" (*flow*), $p = .056$, "I felt imaginative" (*immersion*), $p = .099$, "I felt that I could explore things" (*immersion*), $p = .094$, "I felt skillful" (*competence*), $p = .068$, "I was good at it" (*competence*), $p = .059$, and "I felt successful" (*competence*), $p = .061$. The other dimensions (challenge, tension, affect) showed no significant differences. Levels completed by players using red (7.80) did not significantly differ from players using blue (7.74), $p > 0.05$.

Gender

We wanted to investigate if the previous differences appeared to affect both genders. From Figure 5-21, we can see that the general trend is the same as in Figure 5-20 for both genders (i.e., blue > red across the three measures). However, the effect appears to be weaker in female participants. The effective difference in male participants compared to female participants is 12x larger for *flow*, 3x larger for *immersion*, and 1.3x larger for *competence*. These results are consistent with literature that suggests red is more impactful on men [235, 245].

Text Responses

Using Linguistic Inquiry Word Count (LIWC) 2015 [430], we analyzed text responses of participants' answers to "Describe how you felt about your avatar". LIWC found that negative sentiment was significantly higher for players using the red avatar (6.09) than for players using the blue avatar (3.29), $t(503) = 1.973$, $p < .05$. Positive sentiment in players using the red avatar (8.18) did not significantly differ from players using the blue avatar (8.85), $p > .05$.

Were the colors hard to see?

To determine if the color negatively interacted with the game background, participants were asked “The avatar was hard to see” (1: *Not at all*, to 5: *Extremely*). Both blue participants ($M = 1.36$, $SD = 0.76$) and red participants ($M = 1.30$, $SD = 0.70$) had low scores, suggesting both avatar colors were clearly visible. Scores did not differ between the two conditions, $p > .05$.

Limitations

Color stimuli varies on hue, lightness, and chroma. According to Elliot et. al [145], nearly all existing studies fail to control for these in color manipulations. This makes both interpretation and replication impossible. For example, the majority of research uses hues which the investigators believe are the most ideal representatives. However, the problem is that this almost undoubtedly confounds color properties; for instance, “prototypical red” is more intense than “prototypical yellow”.

The colors in this experiment were selected from the Munsell color system, such that the following criteria were met: (1) the colors are equal in lightness and chroma, (2) the colors do not clash with the game interface, and (3) the colors are accurate on calibrated monitors.

Nonetheless, users each have their own individual monitors, graphic cards, and calibration settings. Not all users will see “exactly” the same color (as in a laboratory setting), but this approach strengthens external validity and reflects realistic applied settings. We do note that our participants were all from the U.S.

Experiment-Specific Discussion (Red vs. Blue)

Our results suggest that avatar color has significant effects on player flow, immersion, and competence. Although we have only investigated two of the colors most prevalent in literature [376], it’s reasonable to hypothesize that other colors may also impact players. For instance, it was found in [268] that different colored *environments* may impact affect. To the best of our knowledge, this is one of the first studies to research the effects of avatar color in a single-player context. These results extend and support work on first-person shooter (FPS)

multiplayer games in which it is hypothesized that blue teams are at a disadvantage because they “see red” [251].

In this study, we found that red had a negative effect on participant flow, immersion, competence, and avatar affect. Biologically, it has been hypothesized that the color red is a distractor signal to humans. Red causes a lower so-called high frequency heart rate variability (HF-HRV), measured via an electrocardiogram (ECG) [143]. These lower levels of HF-HRV are correlated to an increase in worry and anxiety [169, 170, 376].

However, color is context-specific. Although the color red has been found to hinder motivation, performance, and affect in cognitive tasks [146, 190, 244, 271, 314, 329, 374, 376, 493], red has been shown to promote approach-like tendencies when in the context of “dating” [375]. The current investigation used as a setting a computer science learning game, and so it is reasonable to predict that red is hindering. Such effects may translate to changes in academic self-concept [488]. However, were the color red presented in the context of, e.g., a social game ([368, 405], etc.), it’s possible that it’s effects would be less negative.

5.3 Interface Experiments

Experiment listing:

[Feedback Positive vs. Negative vs. Neutral vs. Nothing](#)

[Mini-Game Loss vs. Near-Win vs. Win](#)

[Game Theme Basic vs. Circuit vs. RPG vs. Choice](#)

[Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice](#)

5.3.1 Feedback Positive vs. Negative vs. Neutral vs. Nothing

Previous Experiment: [Red vs. Blue](#)

Category: [Interface Experiments](#)

Next Experiment: [Mini-Game Loss vs. Near-Win vs. Win](#)

Experiment Overview (Feedback Positive vs. Negative vs. Neutral vs. Nothing)

Encouragement (e.g., “You’re doing well”) given at regular intervals improves performance in a variety of sporting domains [13, 128, 205]. This improvement is regardless of the actual performance of participants. However, it has not been studied how this type of encouragement can affect players of video games. In the current study (N = 662), we look at the following encouragement conditions: (1) Positive (e.g., “You’re doing good”), (2) Negative (e.g., “You’re doing badly”), (3) Neutral (e.g., “You’re doing average”), and (4) None. Via the Game Experience Questionnaire (GEQ) [249], participants in the Neutral condition had significantly improved flow, immersion, and affect than participants in the None condition. Moreover, participants in both the Positive and Neutral conditions had the highest overall GEQ ratings. These findings are directly relevant to educational games.

Experiment-Specific Background (Feedback Positive vs. Negative vs. Neutral vs. Nothing)

Simple phrases of encouragement (e.g., “You’re doing well”) delivered at 30-second intervals, significantly improves performance in walking distance [205]. Numerous studies have reproduced similar results in a variety of strength and endurance domains [13, 128, 381, 556]. However, few studies on these types of interventions have been studied in games. O’Rourke et. al found that encouraging the development of a growth mindset, or the belief that intelligence is malleable, increases player perseverance [142, 406]. To the best of our knowledge, however, no studies have attempted to study this simple encouragement inside games.

Encouragement is different from feedback, in that it doesn’t necessarily encode information about performance [303, 384, 444, 478]. Our experiment follows a similar model to previous ones on encouragement [13, 128, 205, 381, 556]. The encouragement is: 1) Always the

Condition	Sentence	Score
Positive	You're doing well	3.10
Positive	Don't give up!	2.44
Positive	You're almost there	2.29
Negative	You work poorly	-3.43
Negative	You're on the wrong track	-2.12
Negative	You're still far away	-1.53
Neutral	You are doing standard work	0.08
Neutral	You're doing average	0.03
Neutral	You're doing typically	0.01

Table 5.14: Example sentences

same valence depending on condition, 2) Speaks to the task at hand and not the learner, and 3) Dispensed at regular time intervals [71, 393, 497]. Our goal is to study how game experience is affected by encouragement, and whether it is positively affected relative to no encouragement at all.

Experiment-Specific Methods (Feedback Positive vs. Negative vs. Neutral vs. Nothing)

Our experiment aims to compare four encouragement conditions: (1) Positive, (2) Negative, (3) Neutral and (4) None. The goal is to see if participants in these conditions have different game performance and game experience as measured by the GEQ. The experiment takes place in *Mazzy* ([278]; Section 3.1).

Creating Sentences

In designing the sentences for each condition, 150 sentences were drafted (50 for each of positive, negative, and neutral conditions). These were developed based on previous encouragement studies [39, 205, 470]. We then recruited 103 U.S. participants to rate the sentences on a scale of -5: *Very Negative* to 5: *Very Positive*. Intraclass correlation on the questions was $ICC = 0.99$ (two-way random, average measures [496]), indicating high agreement.



Figure 5-22: Positive condition.



Figure 5-23: Negative condition.

20 sentences were then randomly selected for each condition. In doing so, each randomly selected positive sentence was matched to the negative sentence with the closest opposite numeric valence score. The average words per sentence did not differ significantly between any of the conditions, $p > .05$. The final average valence scores for the positive sentences was 2.75, for the negative sentences -2.75, and for the neutral sentences 0.00. See Table 5.14 for examples.

Conditions

The four encouragement conditions we tested were:

1. Positive
2. Negative
3. Neutral
4. None

Sentences appeared centered at the bottom of the screen in 28 px font. The words appeared on a 46 px high semi-transparent black bar. Procedure, instructions, gameplay, were all exactly identical across all conditions, only the sentences appearing were different. One randomly chosen sentence was shown at 30 second intervals. Each sentence was displayed for 15 seconds. In the None condition, the black bar was still displayed, but no text was shown. See Figures 5-22, 5-23, 5-24, and 5-25.

Quantitative and Qualitative Measures

For performance, we looked at number of levels completed by players. For measuring game



Figure 5-24: Neutral condition.



Figure 5-25: None condition.

experience, we use the GEQ [249].

Participants

662 participants were recruited through Mechanical Turk. The data set consisted of 51.6% male, and 48.4% female participants. Participants self-identified their races/ethnicities as white (80.5%), black or African American (9%), Chinese (2.3%), Asian Indian (1.2%), Filipino (0.9%), Korean (0.8%), American Indian (0.6%), Japanese (0.5%), and other (4.1%). Participants were between the ages of 18 and 78 ($M = 32.3$, $SD = 9.7$), and were all from the United States. Participants played the game a single time for an average length of 22.9 minutes. Participants were reimbursed \$1.50 to participate in this experiment.

Design

A between-subjects design was used: encouragement valence was the between-subject factor. Participants were randomly assigned to a condition.

Protocol

Prior to starting the game, players were informed that they could exit the game *at any time* via a red button in the corner of the screen. When participants were done playing (either by exiting early, or by finishing all 12 levels), participants returned to the experiment instructions, which then prompted them with the GEQ and then a demographics survey.

Analysis

Player responses were analyzed using multivariate analysis of variance (MANOVA) in SPSS.

The dependent variables were *GEQ items*; and the independent variable was *encouragement (positive, negative, neutral, or none)*. All the dependent variables are continuous variables. The independent variable encouragement (i.e., 0 = positive, 1 = negative, 2 = neutral, 3 = none) was a quadchotomous variable. To detect the significant differences between encouragement conditions, we utilized one-way MANOVA. We also used one-way ANOVA on the variable *levels completed*. These results are reported as significant when $p < 0.05$ (two-tailed). Before running MANOVA, all the variables included in the analyses were checked, and there were 17 outliers detected [236]. These 17 outliers were excluded from further analysis. Prior to running our MANOVA, we checked assumption of homogeneity of variance by the test of Levene's Test of Equality of Error Variances; and the assumption was met by the data ($p > .05$).

Results & Findings (Feedback Positive vs. Negative vs. Neutral vs. Nothing)

Participants in the Neutral condition had significantly improved flow, immersion, and affect than participants in the None condition. Moreover, participants in both the Positive and Neutral conditions had the highest overall GEQ ratings.

Aggregate

The one-way ANOVA found no significant effect of encouragement valence on levels completed, $F(3, 641) = 1.51$, $p = 0.21$ (see Table 5.15).

A MANOVA revealed a statistically significant difference in GEQ responses based on the participant's encouragement valence, $F(126, 1799) = 1.44$, $p < .005$; Wilk's $\lambda = 0.750$, partial $\eta^2 = .09$. See Figure 5-26.

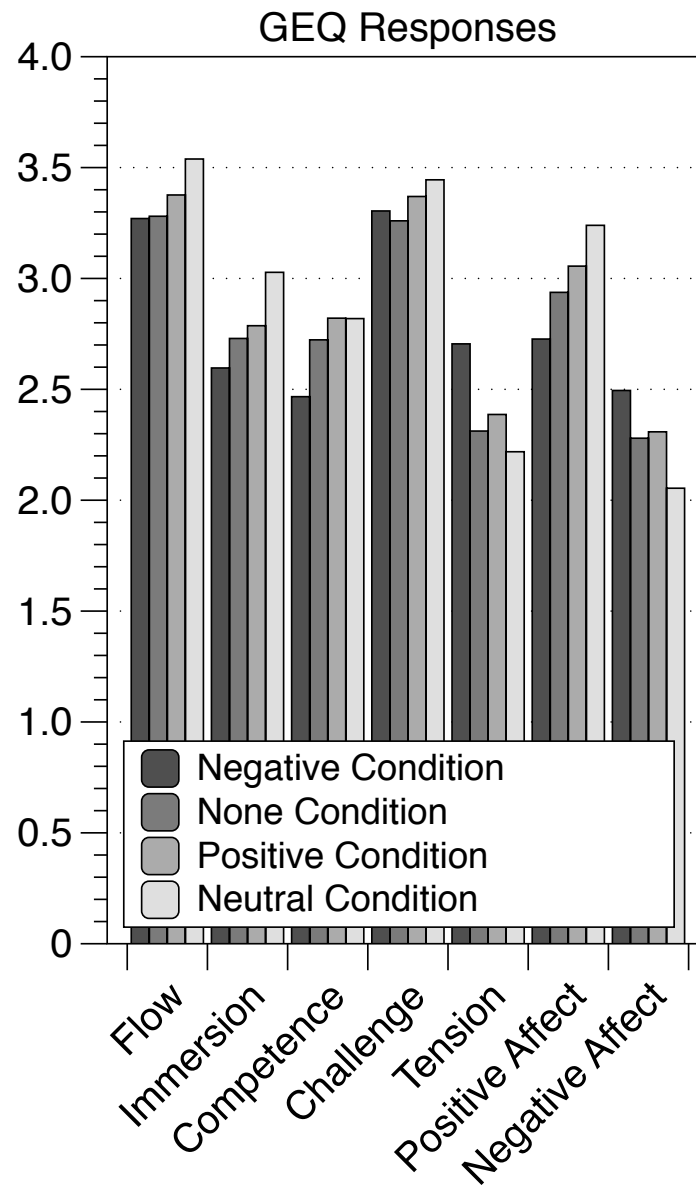


Figure 5-26: Game Experience Questionnaire (GEQ) responses.

Univariate testing found the effect to be significant for the following items:

- **Flow:**

“I felt completely absorbed”, $p < 0.001$.^{ad}

“I was deeply concentrating on the game”, $p < 0.05$.^a

- **Immersion:**

“I was interested in the game’s story”, $p < 0.01$.^a

Valence	N	Mean	SD
Positive	147	7.69	2.79
Negative	161	7.16	2.85
Neutral	151	7.78	2.78
None	186	7.49	2.84

Table 5.15: Overall level completion statistics.

“I felt imaginative”, $p < 0.005$.^{ad}

“I felt that I could explore things”, $p < 0.005$.^a

“I found it impressive”, $p < 0.005$.^a

“It felt like a rich experience”, $p < 0.005$.^a

- **Competence:**

“I felt skillful”, $p < 0.05$.^a

“I felt strong”, $p < 0.05$.^a

“I was good at it”, $p < 0.01$.^b

“I felt successful”, $p < 0.001$.^{abc}

“I was fast at reaching the game’s targets”, $p < 0.01$.^{ab}

- **Challenge:**

“I felt that I was learning”, $p < 0.05$.^a

“I felt stimulated”, $p < 0.005$.^{ab}

“I felt time pressure”, $p < 0.05$.^c

- **Tension:**

“I felt tense”, $p < 0.05$.^{bc}

“I felt restless”, $p < 0.01$.^a

“I felt annoyed”, $p < 0.001$.^{abc}

“I felt irritable”, $p < 0.005$.^{ab}

“I felt frustrated”, $p < 0.05$.^{ac}

“I felt pressured”, $p < 0.005$.^{abc}

- **Positive Affect:**

“I felt content”, $p < 0.001$.^{abc}

“I felt happy”, $p < 0.001$.^a

“I felt good”, $p < 0.001$.^{abc}

“I enjoyed it”, $p < 0.001$.^{abd}

“I thought it was fun”, $p < 0.001$.^{abd}

- **Negative Affect:**

“I thought about other things”, $p < 0.01$.^a

“I found it tiresome”, $p < 0.01$.^{ad}

“I felt bored”, $p < 0.001$.^{abd}

“I was distracted”, $p < 0.05$.^a

“I was bored by the story”, $p < 0.01$.^a

“It gave me a bad mood”, $p < 0.05$.^c

In order to compare the effects of encouragement type on these measures, we additionally calculated posthoc comparisons (Tukey HSD) between all conditions. Superscripts denote cases when Neutral outperforms Negative (^a), Positive outperforms Negative (^b), None outperforms Negative (^c), and Neutral outperforms None (^d). We note that the sheer consistency across all questions indicates an ordering (i.e., Figure 5-26).

Why Does Neutral Outperform Positive?

Participants in the Neutral condition have the highest GEQ ratings (aside from competence). Investigating, we looked at responses to “How did you feel about the [encouragement] text in the game?”. Words most often used to describe the encouragement text in the Positive condition were “encouraging” and “positive”. Participant No. 27 described it as:

“It was encouraging. Made me smile a bit even though I knew I was doing terrible at the game.”

Some participants (13%) found the positive encouragement text less helpful. Participant No. 98 said:

“I liked that the feedback was encouraging, but it seemed “fake” in the sense that no matter what I did, I was going to receive positive feedback. That cheapened it a bit.”

On the other hand, players used words like “indifferent” and “encouraging” to describe the words in the Neutral condition. Participant No. 119 said:

“Sort of helpful. It made me feel a little better knowing I was at least average, when I figured I was totally sucking.”

Participant No. 115 said:

“I felt like it brought me down a little and added a little bit of pressure, yet I could ignore it easily had I wanted too [sic].”

Participants No. 22 and No. 56 suggested that “it [the neutral text] pushed me to work harder” and “it [the neutral text] was humorously neutral”. From these responses, it’s clear that the impacts of encouragement were variable. Virtually all participants found the positive encouragement text to be helpful early on. However, many players that progressed past half-way (Level 7 and onwards) found the text to be “fake”. For these players, the following three things were happening simultaneously: 1) The game was becoming harder, 2) The participants were experiencing frustration, and 3) The positive text combined with the participant’s frustration served only to further increase frustration.

On the other hand, participants in the neutral encouragement condition did not have responses that varied by how far they had progressed in the game. They expressed a level of indifference; a few participants explicitly stated the neutral text had a motivating effect, e.g., to be better than average. Contrary to the participants in the positive encouragement condition, participants in the neutral encouragement condition never felt that the text was fake.

Experiment-Specific Discussion (Feedback Positive vs. Negative vs. Neutral vs. Nothing)

We have seen from our results that players with neutral and positive encouragement had the highest engagement. Our measurement instrument was the GEQ. The GEQ was used for its multiple subscales which assess different components of the player experience, and it is

both a widely used and recognized instrument [402]. Although the GEQ was adequate for measuring engagement, there are a number of viable alternatives [401]. Instruments such as the Big Five Inventory (BFI) [263] could further shed light on the moderating effects of player personality.

Research on Feedback Interventions (FIs) have shown that predicting the effect of any given feedback is contingent on a wide array of factors: personality, feedback type (verbal, etc.), frequency of feedback, task complexity, task novelty, type of task (physical, etc.), etc. [70, 303, 312, 394, 397, 461, 497]. Therefore, researchers should be wary of prescribing general guidelines regarding encouragement.

Keeping in mind the numerous contextual moderators, our results suggest that encouragement can improve game experience. Even in a setting where the encouragement was not directly connected in any way to the gameplay, results showed significant increases in flow, immersion, positive affect, etc. Positive encouragement appeared to benefit players most when the game was easy; those benefits tapered as the game progressively became harder (the results are consistent with work in which insincere praise has a negative effect [229, 440]). Encouragement models that better match player performance, e.g., acknowledging the player's struggles in an encouraging tone, could yield greater benefits.

In this experiment, we have explored the effects of different types of encouragement. We have shown that encouragement (relative to no encouragement) can improve the game experience of players. This is consistent with other work on encouragement [13, 128, 205, 381, 556]. While being mindful of the highly contextual nature of this topic, educational games can consider encouragement as a means to improve game experience. Better engaging learners is one route towards creating more meaningful learning experiences [56].

5.3.2 Mini-Game Loss vs. Near-Win vs. Win

Previous Experiment: [Feedback Positive vs. Negative vs. Neutral vs. Nothing](#)

Category: [Interface Experiments](#)

Next Experiment: [Game Theme Basic vs. Circuit vs. RPG vs. Choice](#)

Experiment Overview (Mini-Game Loss vs. Near-Win vs. Win)

This exploratory study was influenced by various works on how winning/losing can have various impacts on subsequent performance, testosterone, intrinsic motivation, etc. e.g., [31, 43, 61, 68, 72, 92, 112, 119, 147, 155, 156, 267, 366, 447, 448, 486, 502, 532]. It was particularly motivated by Wadhwa and Kim's study which proposed that *just* failing to obtain a reward was more motivating than clearly winning or clearly losing. In their experiments, they found that a near win increased motivation in a variety of contexts [532]. We were interested in whether such a manipulation would positively affect players in our own environment.

Our experiment involved playing *Mazzy*, except immediately before, they played a mini-game. Players are told "While the game loads, you will have a chance to play a mini-game. This game involves clicking on tiles, and uncovering checkmarks or Xs. You have 8 clicks. If you uncover 8 checkmarks, you win the mini game.". In the clear loss condition, players uncover an X in the first click. In the near win condition, players uncover 7/8 checkmarks, and uncover an X on the last click. In the win condition, players uncover all 8 checkmarks.

Overall, we found that the Near Win condition had an insignificant effect. There were some interesting trends, such as longer total play time (~3 minutes more on average), slightly faster level completion time, etc. but statistically these were not significant at our sample size. Being that the study results were inconclusive, we posit that perhaps other factors could have increased the relevance of the mini-game to the player, e.g., offering a tangible monetary reward.

5.3.3 Game Theme Basic vs. Circuit vs. RPG vs. Choice

Previous Experiment: [Mini-Game Loss vs. Near-Win vs. Win](#)

Category: [Interface Experiments](#)

Next Experiment: [Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice](#)

Experiment Overview (Game Theme Basic vs. Circuit vs. RPG vs. Choice)

The results of over twenty-five years of research seem clear: the addition of seductive visual details in video games hinders performance of learners [178, 455, 513]. Yet, countless other research results propose the opposite: that visual embellishments and well-designed ambiguity instead *improve* learners' performance, engagement, and self-efficacy [488, 517, 554]. To shed light on this apparent contradiction, we devised a particular experiment using *game skins* to implement variations in visual themes of a computer game. Game skins are coherent, interchangeable sets of graphical assets that all implement the same underlying game structure while varying the visual appearance (for instance, see Figure 5-27). In particular, we implemented the following four game skins labeled and described as follows: 1) *Generic* theme with no embellishments (simple flat color background), 2) *Fantasy* game theme (forest, snow, and desert adventure backgrounds), 3) *STEM-oriented* theme (computer circuitry background), and 4) *Choice* (the user picks one of the previous three options). Our goal is determining if there are differences in performance, engagement, and self-efficacy between conditions. The upshot is that the generic condition participants had highest performance (levels) and had highest programming self-efficacy—followed by choice, fantasy game setting, circuitry. However, ordering of conditions for engagement was precisely opposite the trend for performance. We conclude by discussing the trade-offs between the two diametrically opposed approaches to game themes and embellishment: instrumental game skins vs. thematic and deliberately embellished game skins.

Experiment-Specific Background (Game Theme Basic vs. Circuit vs. RPG vs. Choice)

One of the largest paradigm shifts in the last thirty years has been movement away from the learning as an acquisition metaphor [490] and instead toward a concept of learning as fundamentally contextually situated [27, 198, 317, 318, 322, 450, 534, 543, 563]. One resultant argument is that people develop deep expertise—*islands of expertise*—that then lead to the formation of overarching themes, abstract enough that they engender further learning both within and outside of the original topic of interest [111, 491]. Given the vast proliferation of educational games, adaptive learning systems, and MOOCs in recent years [564], it is increasingly important to understand the significance how educational content is situated within computer-based learning environments [137, 180, 349, 457, 554], e.g., ranging from STEM-oriented to fantasy settings in educational games. For decades researchers have found that embellishing instruction with fantasy content, improves instructional efficacy, e.g., as in [19, 106, 113, 453, 488]. Games are touted to move beyond the “content fetish” [184] so prevalent in society and to immerse players in an experience where there is *intentional* inefficiency in conveyed content. That is, instead of trying to rush toward “instrumentalized” games [571], it is specifically the embellished ambiguities that create opportunities to explore [174].

Yet, in making this argument we need to account for the fact that this is the opposite of what some researchers in the learning sciences would postulate. The opposing viewpoint holds that that such embellishments would constitute *seductive details* that impede educational efficacy [127, 178, 217, 321, 417, 455, 469, 481, 513]. The *coherence principle* of multimedia learning is a culmination of this line of work. It advises removing any illustration not of fundamental importance to the instructional goal [97, 363].

Here, our goal is to explore and investigate these opposing viewpoints. We consider how three different game skins affect participants’ performance, engagement, and self-efficacy. We find that the more embellished and more ambiguous, game skins thwart performance, but *improve* engagement. Our results suggest that simpler game skins improve performance, but *reduce* engagement. Such a trade-off is particularly important in educational games, in

which both performance *and* engagement are highly desirable to the end goal [56, 225]. We conclude with a reflective discussion on how educators and developers might navigate this dual goal.

Experiment-Specific Methods (Game Theme Basic vs. Circuit vs. RPG vs. Choice)

Our experiment compared the impacts of four game skin conditions: (a) Generic Theme, (b) Fantasy Theme, (c) STEM Theme (circuit board), and (d) User Choice. The goal was to see if participants using different game skins vary in performance, engagement, and self-efficacy. We suspected that (1) the generic skin would have the highest performance, but that (2) the embellished skins would have the highest engagement. The experiment takes place in *Mazzy* ([278]; Section 3.1).

Game Skin Conditions

The four game skin conditions we tested were:

- a. Generic Theme
- b. Fantasy Theme
- c. STEM Theme
- d. User Choice

The generic theme was specifically made to have no embellishments, just flat color. The fantasy theme and STEM (circuitry) theme were heavily embellished in their respective themes (see Figures 5-27, 5-28 and 5-29). A choice condition was included to test if users given choice of game skin have improved performance [20, 35, 106, 115, 149, 164, 215, 266, 295, 355, 422, 472]. This lattermost condition begins with players selecting a game skin—choices always appear in a random order—afterwards all aspects of the game are exactly identical. See Figure 5-30. The player avatar is a blue triangle (Munsell color 7.5PB 5/18)—the avatar color was chosen to minimize interaction effects with game skins. This was later checked post-game, e.g., virtually all players irrespective their condition (given a range of 1: *Strongly Disagree* to 5: *Strongly Agree*) strongly disagreed that the avatar

clashed with the background ($M=1.46$, $SD=0.89$).

Quantitative and Qualitative Measures

Performance was measured as a function of levels completed, number of attempts, and number of hints. Engagement was measured using the Player Experience of Needs Satisfaction (PENS) scale [474] and the Game Experience Questionnaire (GEQ) [249]. Self-Efficacy was measured using the Computer Programming Self-Efficacy Scale (CPSES) [443]. Our instrument was a selected portion of the original CPSES scale. Principal components analysis (PCA) was performed to assess construct validity, with high validity metrics; reliability using Cronbach's alpha was also high, 94.4 percent. See Table D.1.

Participants

1172 participants were recruited through Mechanical Turk (demographics Table 5.16). Participants were reimbursed \$1.50 to participate in this experiment.

Design

A between-subjects design was used: game skin condition was the between-subject factor. Participants were randomly assigned to a condition.

Protocol

Prior to starting the game, players were informed that they could exit the game at *any time* via a red button in the corner of the screen. When participants were done playing (either by exiting early, or by finishing all 12 levels), participants returned to the experiment instructions, which then prompted them with PENS, GEQ, and CPSES, then a demographics survey.

Analysis

Data was analyzed in SPSS using MANOVA. The dependent variables are levels completed, number of attempts, number of hints, and the PENS, GEQ, and CPSES; the independent variable is game skin condition. All the dependent variables are continuous variables.

The independent variable game skin condition (i.e., 0=generic, 1=fantasy, 2=circuitry, 3=choice) was a quadchotomous variable. A MANOVA was run for performance and for each questionnaire. Before running MANOVAs, all the variables included in the analyses were checked. There were univariate outliers and also multivariate outliers, but no outlier was statistically significant so they were retained. One participant was removed for investing minimal effort (0 attempts, 0 levels completed). Prior to running our MANOVAs, we checked both assumption of homogeneity of variance and homogeneity of covariance by the test of Levene's Test of Equality for Error Variances and Box's Test of Equality of Covariance Matrices. Levene's test was met by the data ($p > .05$), but Box's test ($p < .05$) was found untenable. To address this violation, Pillai's Trace was used instead of Wilk's Lambda.

Results & Findings (Game Theme Basic vs. Circuit vs. RPG vs. Choice)

Both embellishment and ambiguity appear to improve engagement but decrease performance. Performance was ordered: generic > choice > fantasy > STEM. Self-efficacy was ordered the same. Engagement was ordered: STEM > fantasy > choice > generic. This was consistent across several measures. The following lists describe these results in terms of performance, self-efficacy, and engagement in fuller detail.

Performance

- Average playtime 21.2 minutes—no notable differences across conditions.
- Overall MANOVA was significant, $p < 0.001$ (Table D.2).
- Univariate tests found all measures to be significant, $p < 0.05$ (descriptives Table D.3, posthocs Table D.4).
- Across all performance measures, performance was consistently ordered: generic > choice > fantasy > STEM (see Figure 5-31).
- Moreover, this effect was found to be true throughout the entire game.

Self-Efficacy

- Overall MANOVA was significant, $p < 0.05$ (Table D.5).
- Univariate tests found eight (of twelve) CPSES questions to be significant, $p < 0.05$ (descriptives Table D.6, posthocs Table D.7).
- On average, similar ordering to performance: generic > choice > fantasy > STEM (see Figure 5-32).

Engagement (GEQ)

- Overall MANOVA was significant, $p < 0.001$ (Table D.8).
- Univariate tests found eighteen GEQ questions to be significant, $p < 0.05$ (descriptives Table D.9, posthocs Table D.10).
- On average, engagement was ordered: STEM > fantasy > choice > generic (see Figure 5-33).

Engagement (PENS)

- Overall MANOVA was significant, $p < 0.001$ (Table D.11).
- Univariate tests found six PENS questions to be significant, $p < 0.05$ (descriptives Table D.12, posthocs Table D.13).
- Consistently, across all questions on autonomy, relatedness, and presence, conditions were ordered: STEM > fantasy > choice > generic (see Figure 5-34).

Choice

- Choice had *no* notable influence on performance, self-efficacy, engagement.
- True even when accounting for the skewed distribution of choices—generic (25%), fantasy (52%), STEM (23%) (descriptives Table D.14).
- One potential explanation is that the choice presented was not very meaningful to participants [150, 164, 292, 462].

Experiment-Specific Discussion (Game Theme Basic vs. Circuit vs. RPG vs. Choice)

Here, we discuss the importance of our findings, why they may have arose, and reflect on how developers and educators might navigate the trade-offs involved in two diametrically opposed approaches to game themes and embellishment.

We first summarize our findings:

- Generic skin condition participants had highest performance
- Generic skin condition participants had highest self-efficacy
- STEM/Fantasy condition participants had highest engagement

Why is this important? Games are clearly becoming ubiquitous—in 2015, the Entertainment Software Association (ESA) estimates that 155 million Americans play video games, 4/5 U.S. households own a device used to play video games, and 42% of Americans play video games regularly (3 hours or more per week) [148]. Moreover, educators are increasingly trying to harness the potential of games for education; embedding content in fantasy settings is quickly becoming pervasive [19, 106, 113, 184, 224, 453, 488]. This approach has also been commercialized, e.g., Classcraft [?], CodeCombat [3], etc. However, developers' knowledge of how such embellishments may affect users in game-like environments is lacking. In the study reported on here, we found that embellishments may have significant effects on user performance, engagement, and programming self-efficacy. The implications are important, e.g., self-efficacy is a strong predictor of women's career choices, especially in regards to STEM fields [45, 65, 410]. Moreover, performance *and* engagement are measures strongly correlated with learning and motivation [56, 225]. Thus, levels of embellishment appear to significantly influence users on a wide variety of crucial constructs.

Why did this happen? We posit that one cause is *seductive details*, which interfere with problem solving abilities in high cognitive load environments [417, 418]. This happens because of three things [217]: *distraction* (taking attention away from the relevant and moving it towards the irrelevant) [481], *disruption* (making it harder to create correct mental schemas) [321], and *diversion* (priming prior knowledge that is unhelpful) [217, 465].

This is well-known in instructional media, where embellishment is known to distract and also create ambiguity (e.g., line sketches vs. 3D graphics) [77, 364, 365, 483]. Yet some researchers argue that embellishment has motivational affordances [174, 192, 408, 417]. Our results provide validity to both arguments—in our study comparing game skins, our results suggest that embellished themes may reduce performance all the while *improving* participant engagement.

What should developers do now? The implications are powerful. That the mere graphical skin of a game can impact users in a variety of important ways means that we can no longer simply assume that embellishing in fantasy is necessarily positive, e.g., [179, 420, 457, 530, 545, 554], nor the inverse. Instead, we advocate to view embellishment holistically. In considering literature from different research fields, multiple, seemingly dichotomous perspectives are reconcilable under the tenet that *no global maximum exists*. Embellishment may affect performance adversely, all the while affecting engagement beneficially.

Our results also suggest another path forward. Developers must invest in compelling and coherent design. We can imagine a type of theme or skin that is elegant, imaginative, and domain-coherent that is a type of *best of both worlds* theme that would lead to high levels of both performance *and* engagement—themes that avoid unnecessary complexity and embellishment while maintaining elegant thematic coherence. In the future, we hope to further untangle the complicated constructs involved in assessing visual themes. Ultimately, such studies may be valuable for educational designers when it comes to creating diverse types of computer-based environments for learning.

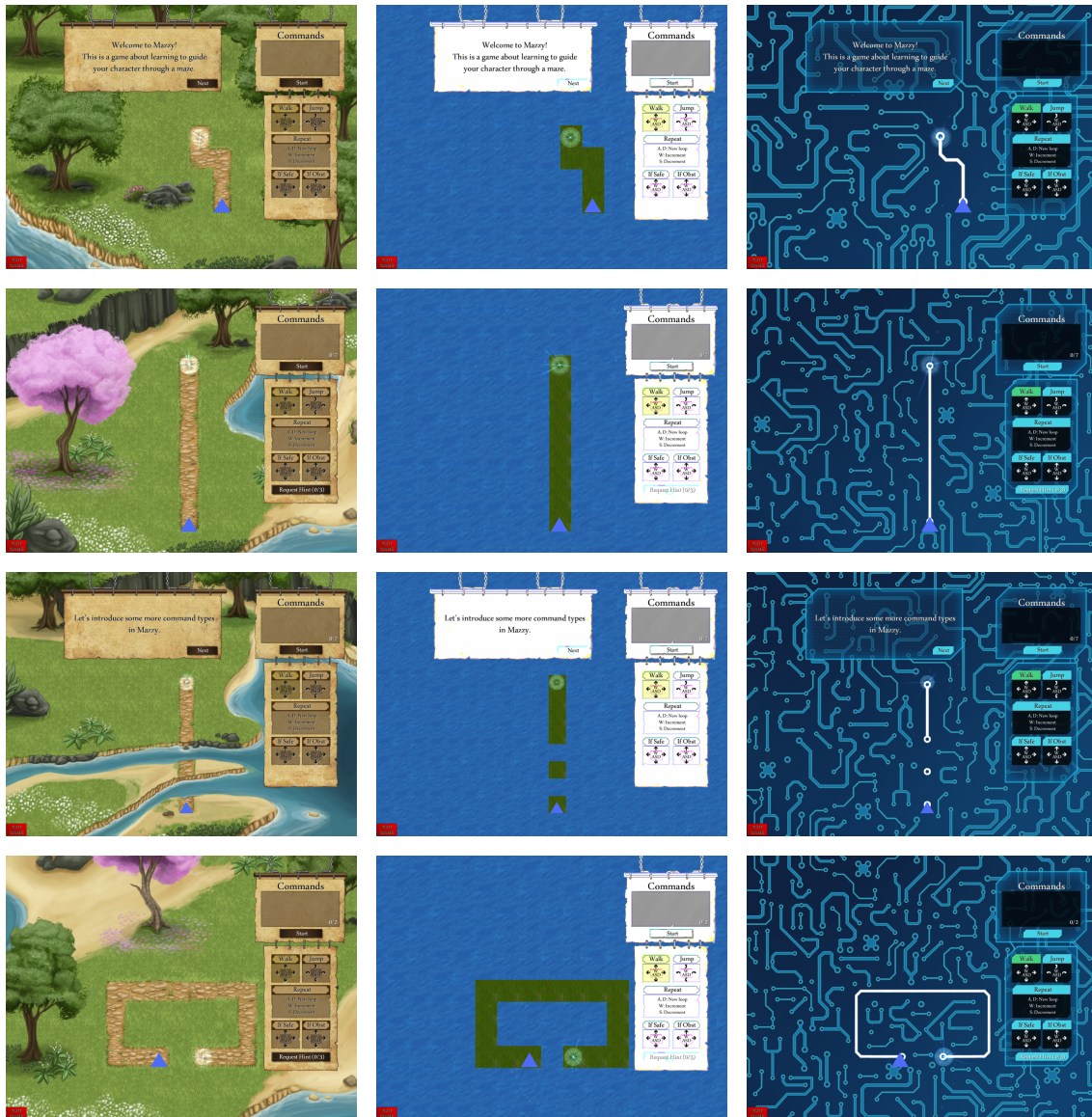


Figure 5-27: Levels 1-4

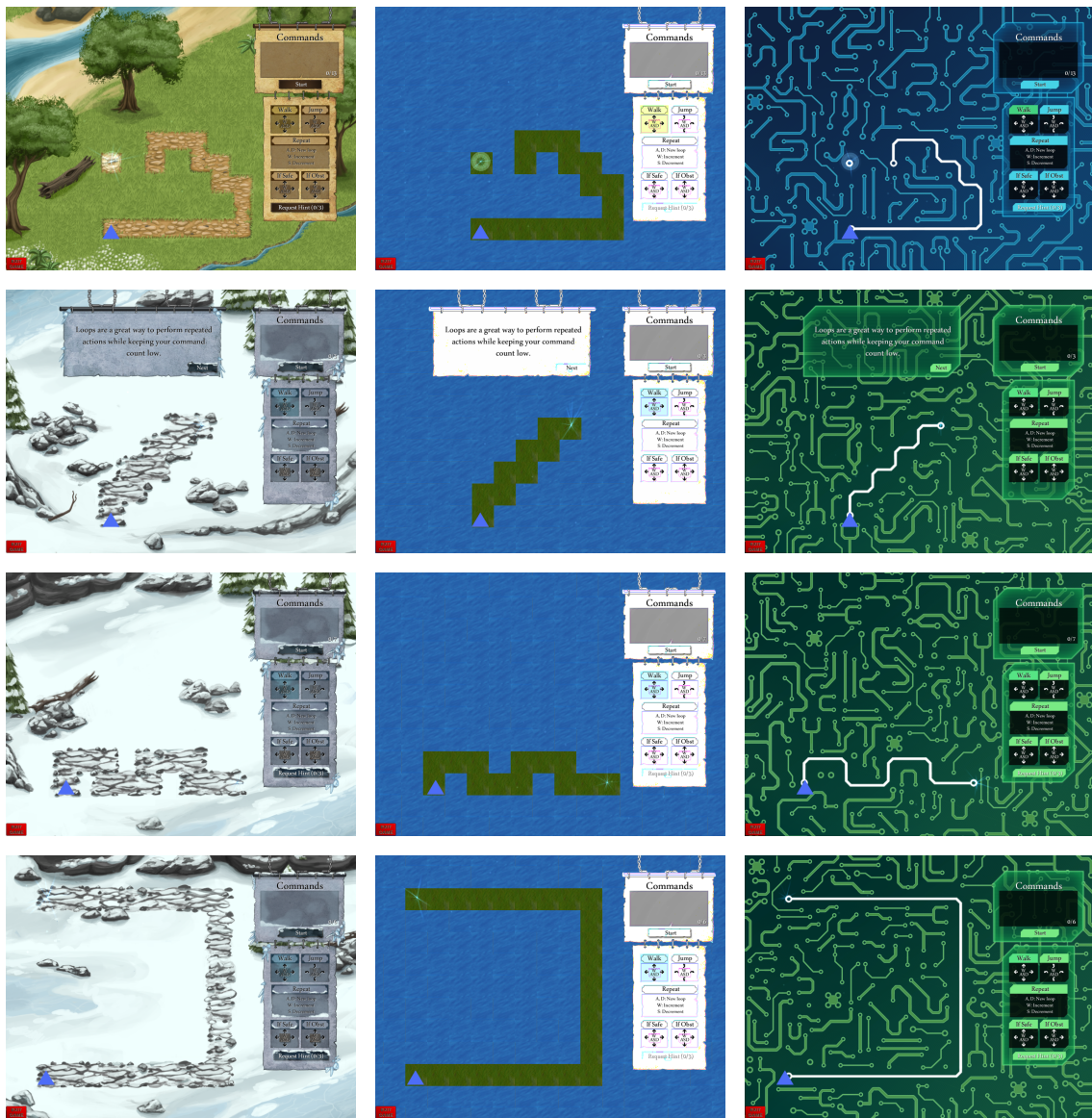


Figure 5-28: Levels 5-8

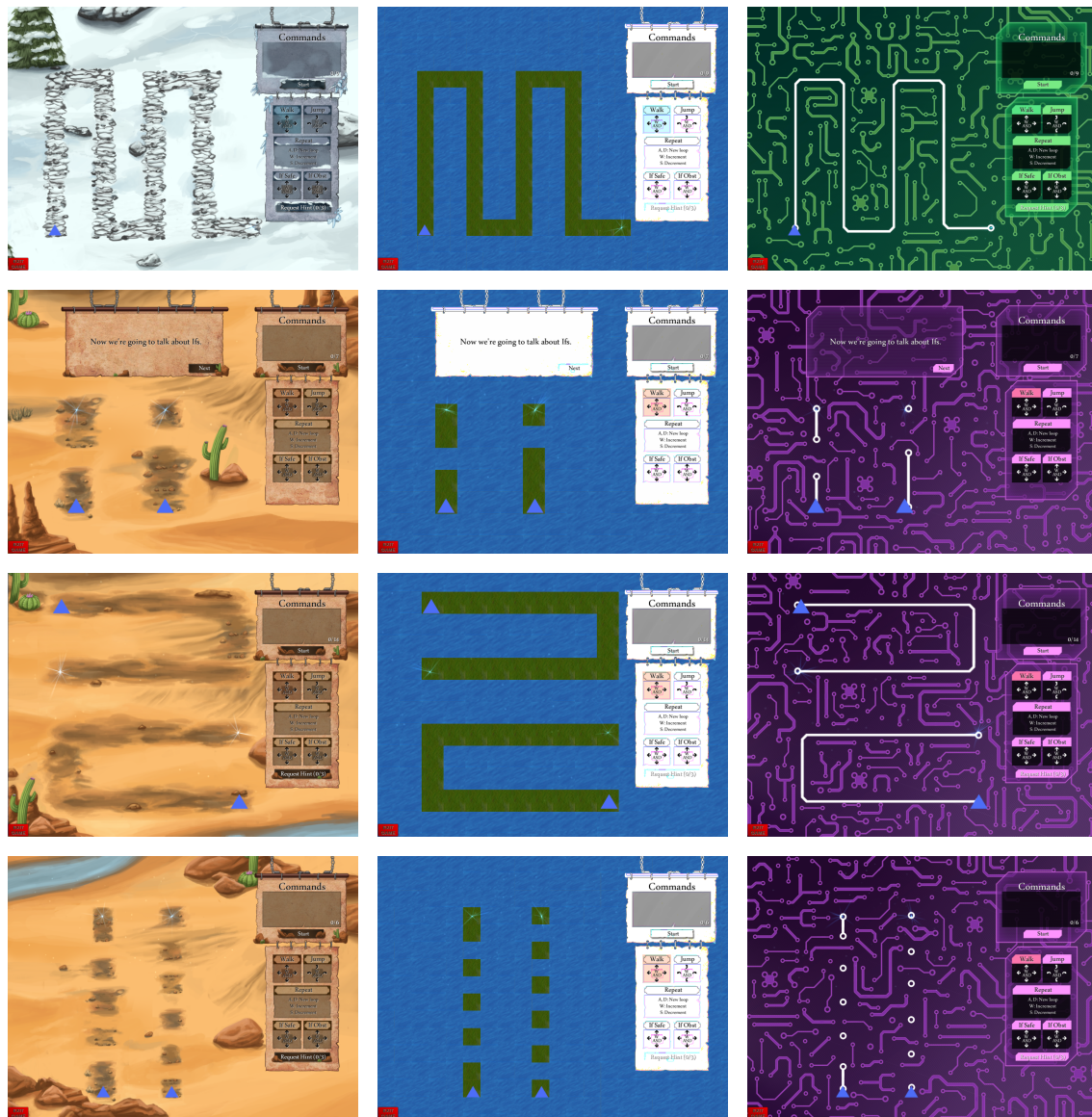


Figure 5-29: Levels 9-12

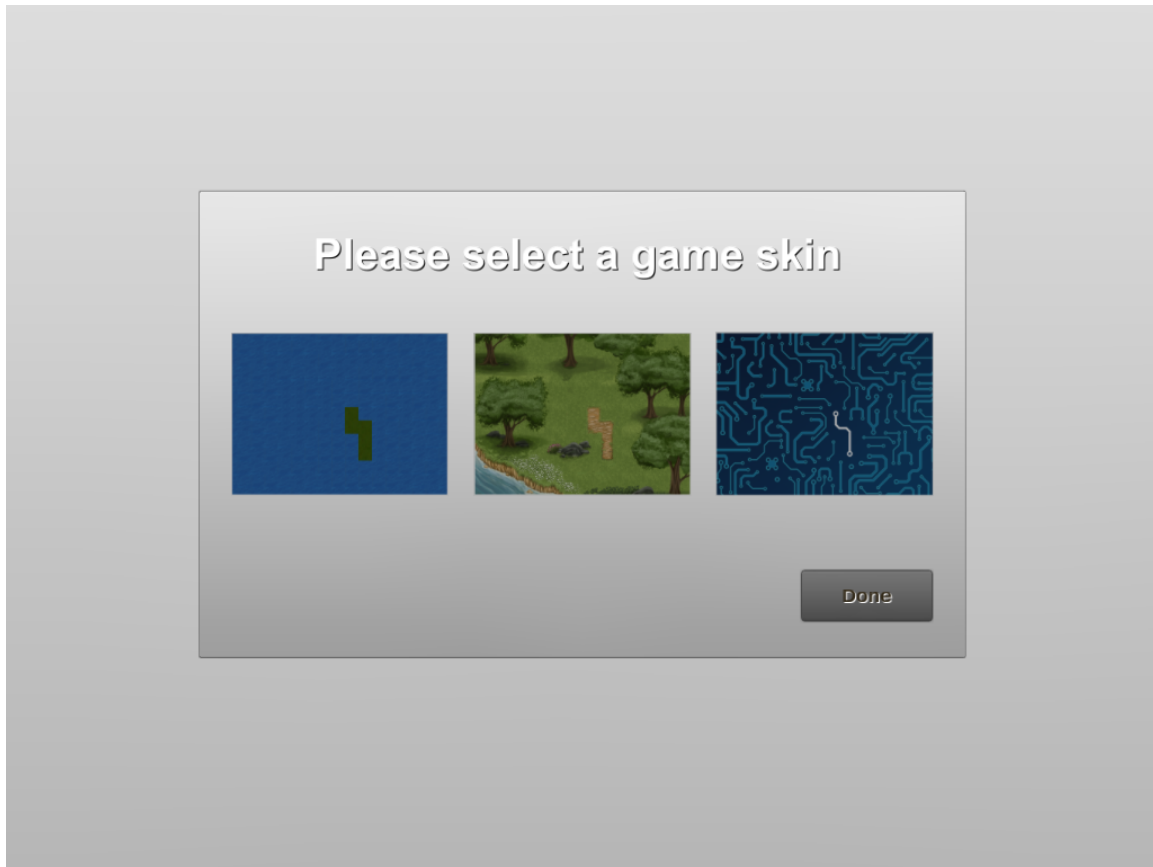


Figure 5-30: Choice Condition

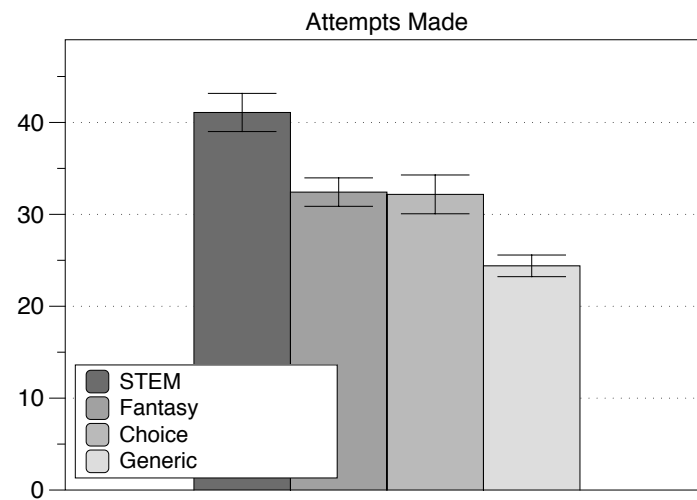
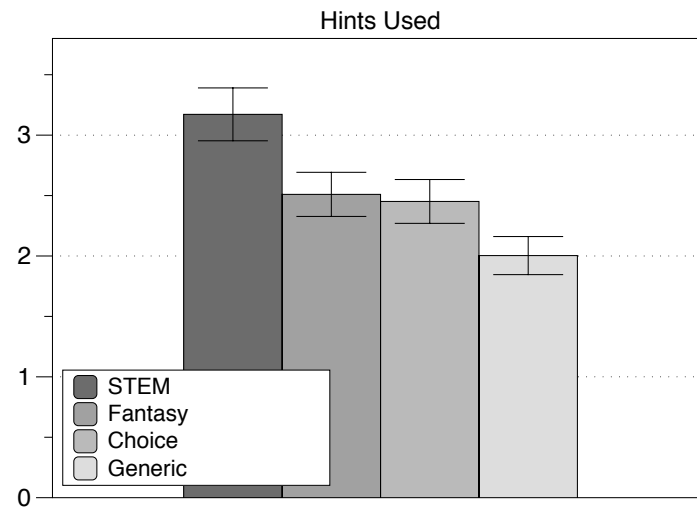
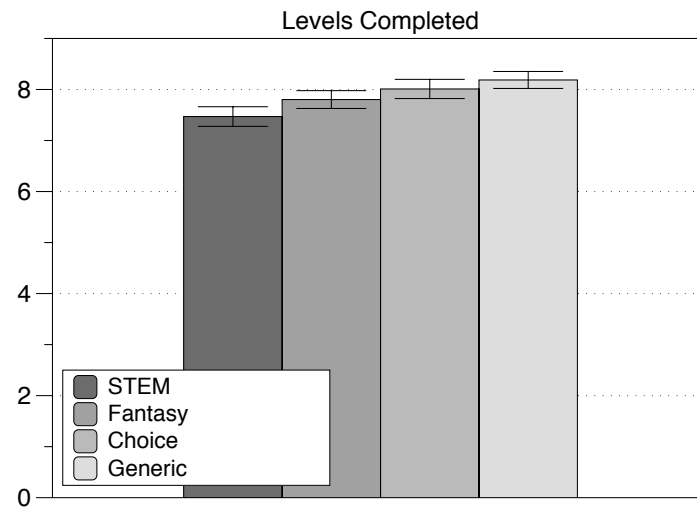


Figure 5-31: Performance—Graphs

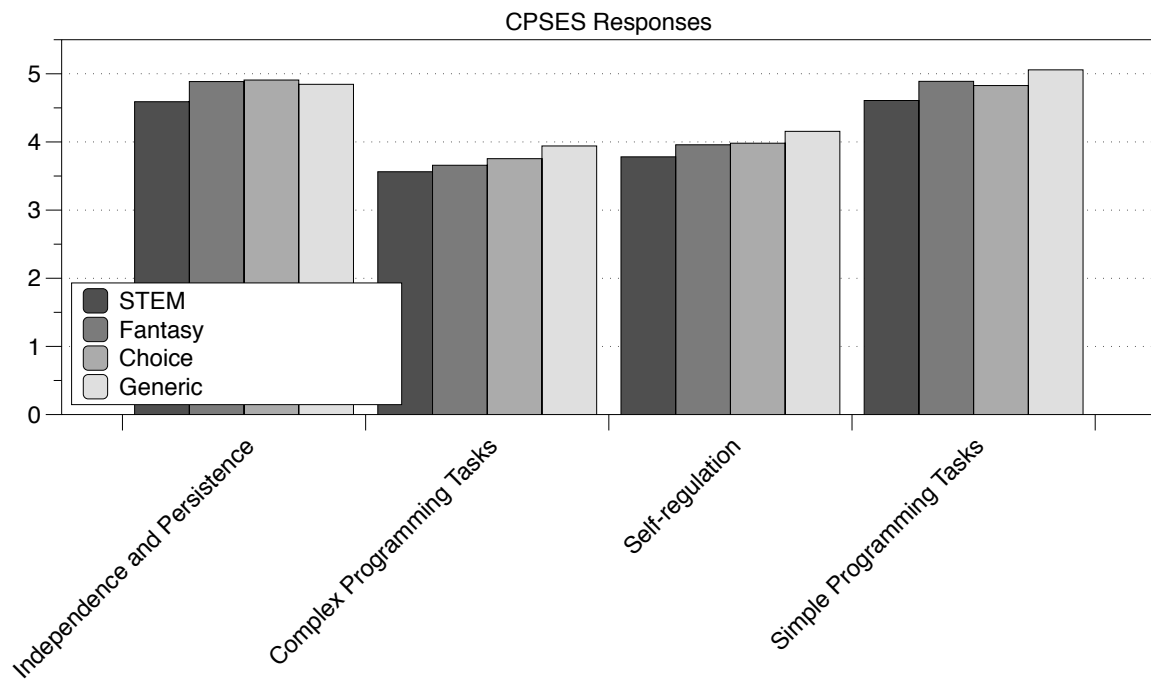


Figure 5-32: Self-Efficacy—Graph

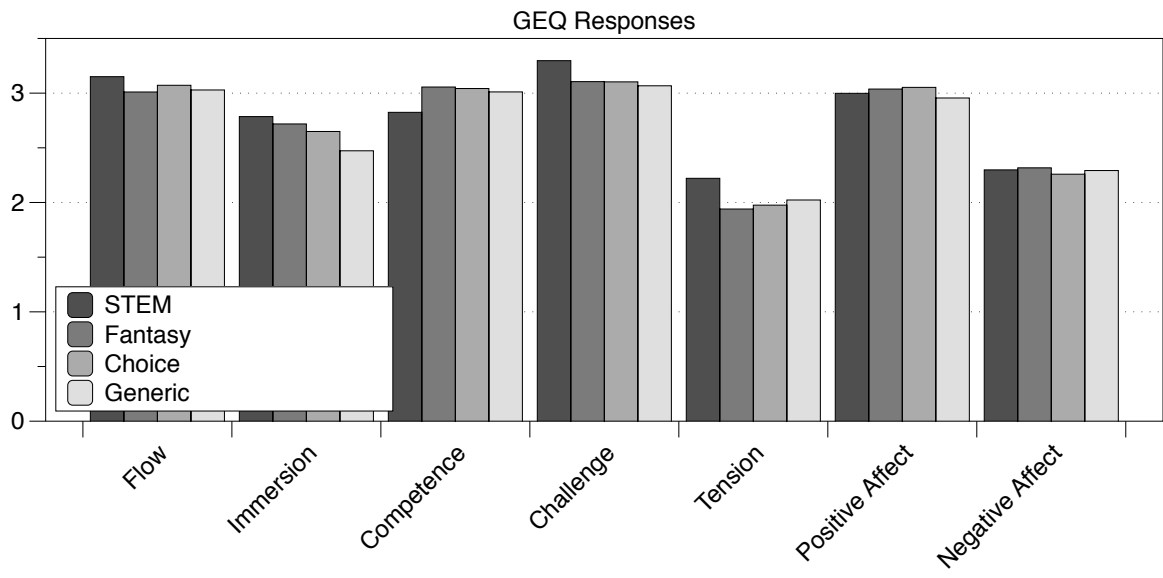


Figure 5-33: GEQ—Graph

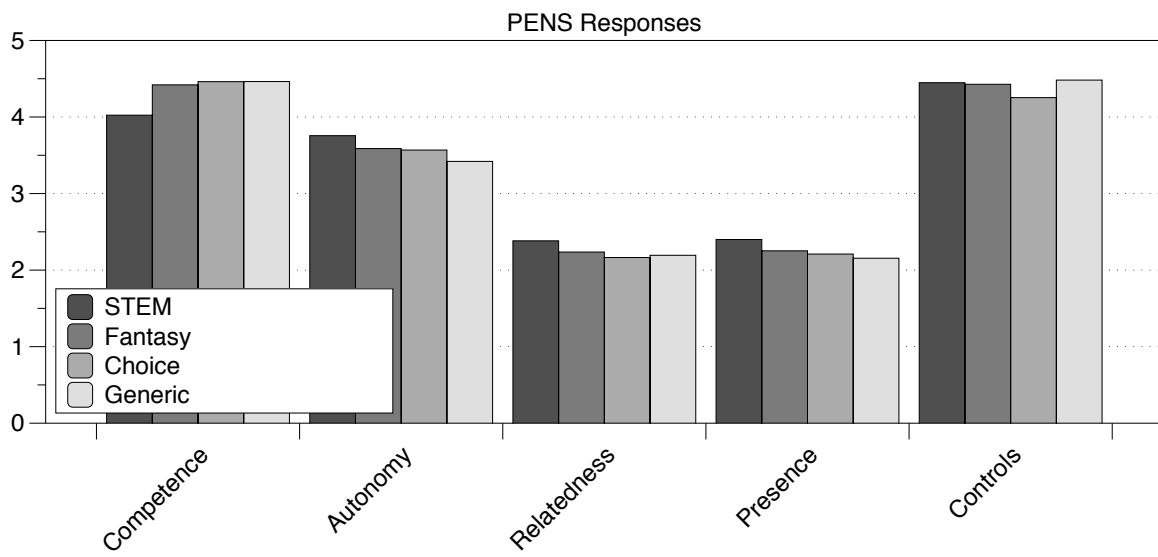


Figure 5-34: PENS—Graph

Characteristic	Category	n	%
Gender	Female	480	41.0
	Male	692	59.0
Age	18-20	73	6.2
	21-30	649	55.4
	31-40	308	26.3
	41-50	99	8.5
	>50	43	3.7
Ethnicity	White	944	80.6
	Black or African American	66	5.6
	Asian Indian	40	3.4
	Chinese	18	1.5
	Korean	8	0.7
	American Indian	9	0.8
	Vietnamese	7	0.6
	Japanese	3	0.3
	Filipino	5	0.4
	Other	72	6.1

1172 participants were recruited through Mechanical Turk. The data set consisted of 692 male, and 480 female participants. Participants self-identified their races/ethnicities as white (944), black or African American (66), Asian Indian (40), Chinese (18), Korean (8), American Indian (9), Vietnamese (7), Japanese (3), Filipino (5), and other (72). Participants were between the ages of 18 and 71 ($M=30.1$, $SD=8.8$), and were all from the United States. Participants were reimbursed \$1.50 to participate in this experiment.

Table 5.16: Demographics

5.3.4 Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice

Previous Experiment: [Game Theme Basic vs. Circuit vs. RPG vs. Choice](#)

Category: [Interface Experiments](#)

Next Experiment: [Badge Type Comparison](#)

Experiment Overview (Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice)

Identical to Game Theme Basic vs. Circuit vs. RPG vs. Choice except all game skins are black and white (see Appendix C). With the goal of exploring the moderating effects of color, the results have not been published as a result of time constraints. At a high-level, the results gleaned from this study suggest that for performance and self-efficacy, ordering of conditions remains the same: generic > fantasy > STEM. However, for engagement, the ordering is: fantasy > generic > STEM. The STEM theme is poorest in terms of both performance *and* engagement. Color appears to play an integral role in the STEM theme.

5.4 Culminating Experiment

5.4.1 Badge Type Comparison

Previous Experiment: [Game Theme Black/White Basic vs. Circuit vs. RPG vs. Choice](#)

Category: [Culminating Experiment](#)

Experiment Overview (Culminating Experiment)

In our study (N=2189), we divided participants into 6 badge conditions: 1) Role model badges (e.g., Einstein), 2) Personal interest badges (e.g., Movies), 3) Achievement badges

(e.g., “Code King”), 4) Choice, 5) Choice with badges always visible, and 6) No badges. Participants played Mazzy, then used the editor to create their own level. Badges promoted avatar identification (personal interest, role model), player experience (achievement, role model), intrinsic motivation (achievement, role model), and self-efficacy (role model) during both the game and the editor. Independent of badges, avatar identification promoted player experience, intrinsic motivation, and self-efficacy. Additionally, avatar identification promoted greater overall time spent in both the game and the editor, and led to significantly higher overall quality of the completed game levels (as rated by 3 independent externally trained QA testers). Our study has implications for the design of badge systems and sheds new light on the effects of avatar identification on play and making.

Experiment-Specific Background (Culminating Experiment)

While the motivational potential of badges has been explored extensively in the last few years [409, 479, 489], badge studies almost exclusively focus on badges as an achievement representation (e.g., [11, 12, 121, 206, 209, 210, 259, 324, 328, 351, 390, 409, 432, 468, 519]). Yet a plethora of studies have demonstrated that being personally associated [4, 138, 552] and having role models [78, 357, 373] is crucial, even affecting career choices in mathematics and CS. As such, we believe that badges representing concepts that are associated with the self, or badges representing relevant role models could enhance motivation, self-efficacy, etc.

Badges also have the potential to enhance avatar identification, one facet of game experience that has been a topic of increasing interest. Avatar identification has been positively correlated to motivation [47, 521], enjoyment [48, 228, 500], long-term participation [298, 512], learning-related outcomes [22, 32, 202, 299], and other player experience outcomes. While avatar identification has been studied in games [47, 499, 512, 521], little is known about how avatar identification impacts constructionist environments. With the rise of constructionist learning in CS [257, 416, 452, 551], and with virtually all environments having a user representation (user profile, avatar, etc.), this is an increasingly important topic. For

example, to what extent are users identified with Logo's turtle, and does it matter?

In this study, we had 4 research questions:

RQ1: Do badges improve identification, need satisfaction, intrinsic motivation, and programming self-efficacy in the CS programming game and the subsequent game-making task?

RQ2: Does avatar identification improve need satisfaction, intrinsic motivation, and programming self-efficacy in the CS programming game and the subsequent game-making task?

RQ3: Does avatar identification translate into higher motivated behavior (time spent, etc.)?

RQ4: Does avatar identification improve created game levels?

We ran a between-subjects study on Amazon Mechanical Turk. Participants played a CS programming game, then used an editor to create their own level. Participants were divided into several badge conditions, including role model badges (e.g., Einstein), personal interest badges (e.g., Movies), and achievement badges (e.g., "Code King"). Independently from badges, we used hierarchical regression with avatar identification as the predictor.

Badges contributed to greater avatar identification (personal interest, role model), player experience (achievement, role model), intrinsic motivation (achievement, role model), and self-efficacy (role model).

Avatar identification consistently improved player experience, intrinsic motivation, and self-efficacy. Avatar identification promoted greater overall time spent in both the game and the editor, and led to significantly higher overall quality of the completed game levels (as rated by 3 independent externally trained QA testers).

To the best of our knowledge, this is the first study to look at badges that represent completely alternative types. To the best of our knowledge, this is also the first study to look at avatar identification in a making context. Our study has implications for the design of badge systems and sheds new light on the effects of avatar identification on play and making.

Badges

What Is a Badge?

Badges summarize achievement and signal accomplishment [187]. Badges are used in games (e.g., Xbox 360 [259]), commerce (e.g., eBay, Amazon), education (e.g., Khan Academy), as physical status icons (e.g., ribbons, medals, trophies), and countless digital applications (e.g., foursquare, Nike+). Educationally, badges can motivate [11, 153, 350], scaffold [11, 260, 269], and credential [8, 345, 392]. In this paper, we are primarily interested in the motivational potential of badges.

In defining badges, it is impossible not to also describe the larger discourse surrounding *gamification*. Gamification, or the use of game design elements in non-game contexts [125], has been studied ubiquitously in education [133, 191, 369], online communities [50, 172, 515], health [241, 463, 507], and innumerable other domains [489]. Gamification has been applied to schools (e.g., Quest to Learn [480]), crowdsourced science (e.g., Foldit [297], Galaxy Zoo [338]), and other domains.

Yet gamification has been contentious [59, 60, 122, 123, 451], critics have argued that its approaches involve “taking the thing that is least essential to games and representing it as the core of the experience” [459]. A centerpiece of this discourse is the notion that extrinsic rewards—e.g., external rewards such as points or in-game currency—can undermine intrinsic motivation [117], i.e., engaging in a behavior because it is satisfying in and of itself [473]. However, meta-analytic reviews to date have not supported this argument [129, 152, 489]. Moreover, this debate is generally centered around non-game contexts (rewards such as badges, leaderboards, points, etc. are viewed as essential game components [125]). Our aim here is to study different types of badges in an educational game.

What Are the Effects of Badges?

Researchers have found that badges increase user activity, e.g., posting trade proposals in a trading service [209, 210] and affect behavior [11, 206, 351], e.g., Q/A behavior on Stack Overflow [11, 351]. Badges have been shown to increase student contributions [121],

promote higher grades [133], enjoyability [162], and can affect the behavior of students even when badges have no impact on grading [206]. Researchers caution, however, that badges can have a negative effect on written work [133] and foster undesirable behavior [162].

Researchers have also found that badges increase forum engagement in MOOCs [12] and productivity of Wikipedia contributors by 60% [454]. The motivational effects of badges can vary by person, activity, and badge [5, 62, 160, 434]. Typically, badges are awarded for fulfilling criteria such as accomplishment [133], participation [11], carefulness [206], and behavior change [7, 335]. Isolating badge effects is often challenging, as many researchers leverage multiple game elements [211, 324, 489]. However, a systematic review in education of game elements, including badges, have found mostly encouraging results [129].

What Types of Badges Have Been Studied?

Badges are widely considered synonymous to achievements [11, 15, 544, 570] and are depicted as achievements across virtually all research studies [11, 12, 121, 206, 209, 210, 259, 324, 328, 351, 390, 409, 432, 468, 519]. While badges have also been discussed as a mechanism for feedback [458], guidance [536], etc., these have been secondary. Most commonly, badges are awarded for performance or completion [51]—for example, “Faster than Lightning” [5], “3D Expert” [519], “Y U No Make Mistakes?” [206], “Almost finished!” [216], “Curious” [351], “Question answerer” [121], etc.

Researchers have sought to categorize badge types from several mainstream video games including Mass Effect, Grand Theft Auto 4, and World of Warcraft: Tutorial, Completion, Collection, Virtuosity, Hard mode, Special Play Style, Veteran, Loyalty, Curiosity, Luck, Mini-Game, Multi-Player, Paragon, and Fandom [385]. In all cases, however, badges depict a type of achievement. Other researchers note that badges may vary in their signifier (the actual visual badge), completion logic (the conditional logic required), and reward [208]. In this study, we are interested in comparing typical achievement badges with badges awarded for the same reasons, but that represent other things, e.g., a personal interest.

Personal Interests

Researchers have argued that students do not generally value science as personally relevant [10, 320], and that games should utilize the affordances for personal relevance [165]. Researchers argue that by incorporating personally relevant items, games can help students both develop schematic knowledge while building a personal connection [165]—see [4, 10, 106, 138, 165, 552] for similar arguments. Researchers suggest that badges that utilize personal identity would: 1) build on identity that is more firmly established, and 2) strengthen positive associations between learning and the student’s identity [4]. One badge type we study is personal interests.

Role Models

Role model—a reference that occupies a desirable standing—was coined by Robert K. Merton [82]. Role models can increase academic performance [78, 357, 373], even affecting career choices in mathematics and CS. Three factors increase role models’ effectiveness: 1) Perception of competence [358], 2) Perception of being ingroup, e.g., shared attributes like gender and race [340, 357], 3) Perception of success [78]. In a CS programming game, participants using role model avatars, e.g., Einstein, have significant increases in flow, immersion, etc. [277, 279, 285]. In a study on group brainstorming using a virtual environment, participants that used scientist-like avatars had more original ideas [203]. Towards understanding if benefits from using role model avatars can be applied in other forms, we investigate role model badges.

Avatar Identification

What Is Avatar Identification?

Avatar identification—sometimes called “player-avatar identification” [326], “character identification” [500], or “avatar-self connection” [262]—is a *temporary alteration of media users’ self-concept through adoption of perceived characteristics of a media person* [95]. Building upon Cohen’s work [101], Klimmt et al. argue that during exposure to a video game, users become one with their character [95]. Extensive work exists on identification

with television characters [102, 237, 238, 506]. However, one important difference is that video games emphasize agency [175]. The active participation in video games is argued to override the distance between user and character [228, 302]. Avatar identification is strongly moderated by similarity, such as demographics, experiences, etc. [101] and other variables, e.g., game type [518]. Other work has shown that users realize their “ideal selves” through games [44], both physically and psychologically [140]. Research suggests that we slowly become more congruent with our virtual identities over time [140, 445, 560, 562].

What Are the Effects of Avatar Identification?

Avatar identification can improve game enjoyment [48, 325, 399, 518], health outcomes [300], intrinsic motivation [47, 521], flow [500], exercise motivation [325, 531], and trust in others [298]. Avatar identification can also reduce self-discrepancy (the distance between one’s actual and ideal selves [234]) [44, 334], improve self-esteem [538], game loyalty [512], learning interest [21], game appreciation [64], game motivation [527], decrease deceptive behavior [240], increase willingness to purchase game items [419, 569], and has been associated with aggression [311], addiction [499], depression [44, 370], and increased persuasiveness of messages [391].

In education, there is a long history of work in avatars and pedagogical agents (i.e., virtual pedagogical agents, teaching agents, etc.). In particular, a large body of work has shown that avatars and agents that share users’ external characteristics (e.g., age, gender, race, clothing, etc.) are more influential and are linked to better learning outcomes [22, 32, 202, 299]—i.e., similarities [518, 526]. This is posited to be a result of similarity-attraction, the theory that people are attracted to similar others [79, 256]. Functional neuroimaging has found that perceived similarity is an important factor in a person’s ability to simulate the internal state of another person [378]. Mobbs et al. found that when a participant watched a game show contestant with high perceived similarity, the participant experienced significant increases in both subjective and neural responses to vicarious reward [380]. Furthermore, work has suggested that what is experienced by an avatar is also experienced by its user [83, 386, 546, 558]. This effect is more powerful via avatars that we identify with [141, 528],

identification being positively correlated to such factors as representation of emotions and intent [212], physical resemblance [346], and avatar customization [520].

For instance, Birk, Atkins, Bowey and Mandryk, divided participants into two groups, one that customized their avatar and another that watched a video of their avatar being customized. Those participants that customized their avatar had increased identification. Furthermore, participants' identification with their avatars significantly predicted various measures related to engagement such as affect, immersion, and amount of time playing [47]. While similarity plays a key role in enhancing avatar identification [518, 526], creating ideal versions of ourselves—sometimes referred to as wishful identification [237, 238]—can also be of value [441]. But this same discrepancy between the actual and ideal self is predictive of negative health outcomes, such as depression and anxiety [193, 234].

Through the clear predictive capacity of avatar identification, we seek to study avatar identification in a making context and its relationship to various facets. For instance, with highly active interest in making [67, 118, 274, 276], we have yet to scratch the surface of topics like identification [274]. Does identification with an on-screen object—a profile picture, a Mii, a turtle, an avatar, etc.—enhance users' making experience? Towards investigating this question, we investigate avatar identification in the context of making.

The Game

The first phase of our experiment takes place in *Mazzy* ([278]; Section 3.1).

The Editor

The second phase of the experiment takes place in an editor [276]. At a high-level, the editor allows players to create their own *Mazzy* game levels. Each map consists of a grid of tiles, each of which can be textured separately and modified logically to be a safe or unsafe tile for the player to step on. The maps can be any size (from 1x1 to, e.g., 100x100). Basic functionalities of the editor include: manipulating the view, creating assets that can be translated, rotated, rescaled, searching for images via a built-in image search that interfaces with Microsoft Bing, and testing maps by playing them. Although the editor typically

provides pre-loaded images for use, in this study none were provided, i.e., all images were searched for.

Experiment-Specific Methods (Culminating Experiment)

Creating Badges

Our goal in this step was to create a set of achievement, role model, and personal interest badges. In order to do so, we populated the initial badge list through crowdsourcing. We validated the badges to ensure that they were adequate for use in our experiment also through crowdsourcing. Since our actual study took place on Amazon Mechanical Turk (AMT), these validation studies also used AMT participants.

Initial Population of Interests and Scientists

100 participants were asked to list: 1) 10 of their personal interests, and 2) 10 scientist role models. No restrictions were made as to the types of interests or scientists.

Interest Corrections

The 1000 interests generated from the preceding step then underwent light corrections. During these corrections, similar interests were renamed, e.g., “Video Gaming” and “videogames” were renamed to “Video Games”. Typos were also corrected.

Interest Categorization

113 participants categorized the 1000 interests. Users were instructed to write a high-level category for each interest, e.g., for “Playing Drums”, they might write “Playing an Instrument” or “Music”.

Popularity Ranking of Interests and Scientists

241 unique scientists, 470 unique interests, and 1005 unique categories of interests were ranked by how frequently different users had mentioned them.

Interests: Food Interest, Family Interest, Movies Interest, Technology Interest, Music Interest, Reading Interest, Nature Interest, Games Interest, Television Interest, Comedy Interest, Animals Interest, Traveling Interest, Health Interest, Cooking Interest, Science Interest, Internet Interest, Fun Interest, Life Interest, Knowledge Interest, Money Interest, Intelligence Interest, Self-Improvement Interest, Creativity Interest, Pets Interest, Fiction Interest, Universe Interest, Exploring Interest, Creating Interest, Culture Interest, Society Interest

Scientists: Albert Einstein, Isaac Newton, Stephen Hawking, Nikola Tesla, Marie Curie, Charles Darwin, Galileo Galilei, Thomas Edison, Carl Sagan, Neil deGrasse Tyson, Alexander Graham Bell, Louis Pasteur, Leonardo da Vinci, Niels Bohr, Jane Goodall, Aristotle, Nicolaus Copernicus, Bill Nye, Gregor Mendel, Archimedes, Michio Kaku, George Washington Carver, Rosalind Franklin, Rachel Carson, Mae Jemison, Lise Meitner, Mary Somerville, Ibn al-Haytham, C. V. Raman, Ada Lovelace

Achievements: Baby Steps, Setting Sail, A New World, Spring in Your Step, Scenic Route, Straight Runner, Taking Off, Water Crosser, Zero Gravity, Step Saver, Creative Solution, Fly By, Labyrinth Master, Perilous Pathways, Lucky Leaper, Loophole, Tip of the Iceberg, Round Trip Flight, Fleet Footed, Gaining Traction, Seventh Heaven, Repeat Runner, Sub-Orbital, Making Camp, Mountain Guide, Prodigy Walker, Beep Boop Beep, What a Breeze, Look Out, World Explorer, Conditional Victory, Code Warrior, Extreme Conditions, A New Dawn, Code King, 100% Worth

Table 5.17: Final set of interests, scientists, and achievements

Final Scientists

For the final set of 30 scientists, we selected the top 20 most often mentioned scientists, and the remaining 10 were researcher-curated from scientists that were mentioned by at least 2 different users, and were diverse in gender and race. While every scientist mentioned had a record of success (i.e., competence [78, 358]), this curation was done to ensure more coverage and inclusiveness (i.e., in-group potential [340, 357]). See Table 5.17.

Final Interests

209 participants rated 100 interests (50 most mentioned interests, and categories). Each rating was done with the format “I have an active interest in _____” on a scale of 1: *Strongly Disagree* to 7: *Strongly Agree*. For the final set, we selected the 15 most highly rated interests, and 15 most highly rated categories. We additionally checked that the averages of this final set of 30 did not differ by gender or race to ensure that personal interests were widely represented, $p > .05$. See Table 5.17.

Finding Badge Images

We used Google to find badge images for each personal interest (searching for “_____ - _ Icon”) and for each scientist (searching the name directly). We then took the very first search result and two more results at random from the first 10.

Processing Badge Images

All potential badge images were cropped to be square. All images were converted to 8-bit black and white to normalize color [146, 244, 287, 376].

Rating Badge Images

107 AMT users then ranked the 3 potential badge images for each personal interest and scientist. For personal interests, users were asked to rank the images from best to worst in their representation of the interest. For scientists, users ranked the images from best to worst in their representation of the scientist. If they did not recognize a scientist, they were asked to find out more about the scientist (a link to the scientist’s Wikipedia page was provided). Both question order and image order were randomized. Final rankings did not differ by gender or race, $p > .05$. Intraclass correlation on the rankings was $ICC = 0.94$ (two-way random, average measures [496]), indicating high agreement.

Finding Achievement Badge Images

With a graphic design artist, we created or found 5 potential achievement images for each game level. These were created to look like an achievement and match the game level context.

Ranking Achievement Badge Images

122 participants played Mazzy. After each level, users ranked the 5 potential badge images for that level. Users provided their own captions for the top three badges they ranked. Rankings did not significantly differ by gender or race, $p > .05$. Intraclass correlation on the rankings was $ICC = 0.94$ (two-way random, average measures), indicating high agreement.

The three highest ranked images for each level were kept.

Achievement Badge Captions

Captions were selected from participants' responses. Captions were selected to match the badge image and the game level context. See Table 5.17.

Number of Caption Words

We performed an ANOVA on number of caption words by badge type, and found no significant effect, $F(2, 93) = 1.42, p = 0.25$.

Badge Image Stylization

To ensure that all badges had a uniform design, we first removed transparency and pure white backgrounds replaced by a neutral gray (rgb[212, 212, 212]). We then used a stylization filter called the “Photocopy” filter in Adobe Photoshop CS5. This served both to provide a uniform design and also to normalize average pixel intensity. The average pixel intensity across badge types did not differ, $F(2, 93) = 1.39, p = 0.25$.

Badge Image Representativeness

100 participants rated the stylized badge images. They were asked “The picture on the right represents the person, icon, or illustration depicted in the picture on the left.” on a scale of 1: *Strongly Disagree* to 7: *Strongly Agree*. A one-way ANOVA found no significant effect of badge type on the representativeness of the stylized images, $F(2, 93) = 1.39, p = 0.26$.

Final Badges

All badge images were then given a circular frame and a ribbon with caption text. This was done by a graphic designer who was instructed to prioritize uniformity, e.g., spacing in the text, etc. All final badges were vetted.

Final Badge-Likeness

103 participants were presented with each final badge and the *original* image—before any

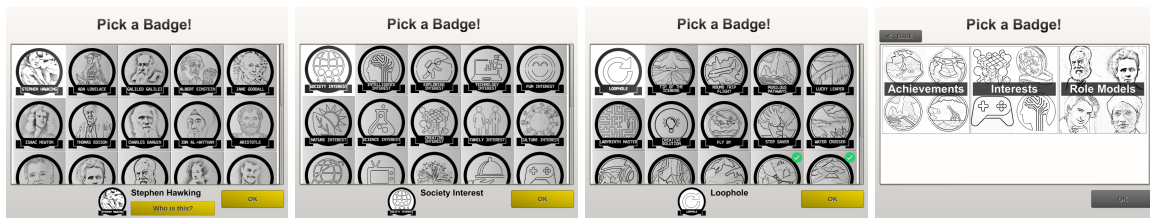


Figure 5-35: Conditions: a) Role Model, b) Personal Interest, c) Achievement, and d) Choice.

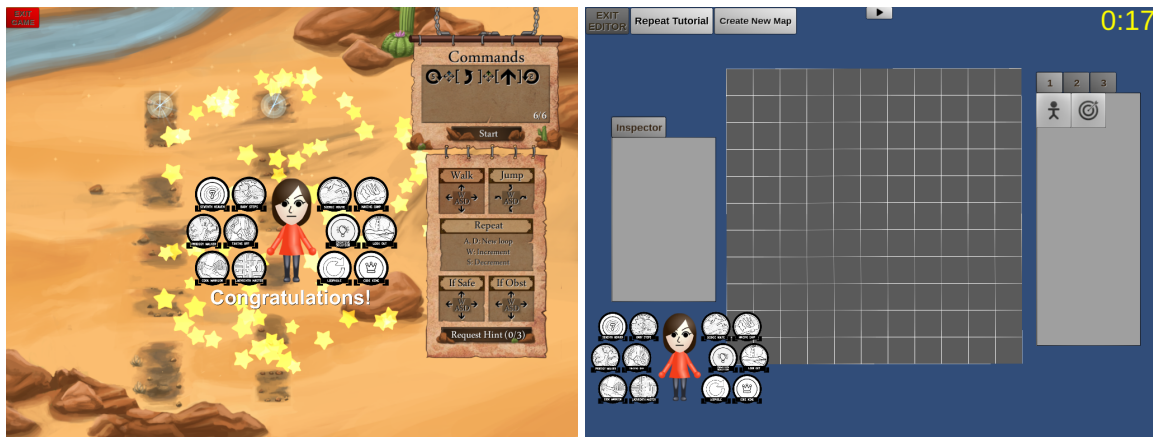


Figure 5-36: Badges as they appear: a) In-game, and b) In-editor.

changes. They were asked “The image on the right is a good badge representation of the image on the left” on a scale of 1: *Strongly Disagree* to 7: *Strongly Agree*. A one-way ANOVA found no significant effect of badge type on goodness of badge representation, $F(2, 93) = 0.56, p = 0.57$.

Final Badge Aptness to Game

110 participants watched a gameplay video of Mazzy. For each final badge, they were asked: “This is a suitable badge for Mazzy” on a scale of 1: *Strongly Disagree* to 7: *Strongly Agree*. A one-way ANOVA found no significant effect of badge type on suitability of badge, $F(2, 93) = 0.47, p = 0.63$.

Conditions

The six badge conditions we tested were:

1. Role model
2. Personal interest
3. Achievement
4. Choice
5. Choice with badges always visible
6. No Badges

Badges are always awarded at the end of each level of the game. See Figure 5-35. **All badge conditions except *Choice always visible*:** During the game, badges are visible when the player completes a level briefly (1.5 seconds), and briefly again after having chosen a badge (1.5 seconds). See Figure 5-36. **All badge conditions except *Choice always visible*:** During the editor, badges are always visible in the bottom left. See Figure 5-36. Badges always appear alongside the avatar. In the editor, the user avatar and badges are positioned such that a distance of 33 pixels exists between the leftmost edge of the window and the nearest badge or avatar pixel. **Choice always visible:** In the *Choice always visible* condition, badges are shown at the bottom of the screen at all times. All other aspects are identical. **No Badges:** In the *No Badges* condition, completing a level shows the Congratulations screen for 3 seconds (with only the player avatar).

Condition specifics: **Role model condition:** Users select from 30 scientists, and can hover over a “Who is this?” button to see a 3-sentence summary taken verbatim from Wikipedia. This is in the form of a semi-transparent black overlay that appears at the bottom. **Personal interest condition:** Users select from 30 interests. **Achievement condition:** Users unlock three achievement badges per level completed. After completing a level, they select from the newly unlocked badges as well as any previously not chosen badges. Newly unlocked achievements appear at the head of the list. There are 36 achievement badges. **Choice condition:** Users choose from all three badge types. Each badge type works the same way as its individual condition. Badge types are presented in randomized ordering, and background images are taken randomly from each individual subset. See Figure 5-35. **Choice always visible condition:** Identical to *Choice* except for display of badges. **No Badges condition:** No badges. **All badge conditions:** Previously chosen badges are marked with a green

checkmark, and cannot be chosen a second time. Badges are randomized in their ordering for each user.

A Note on Choice

We implemented choice in all badge conditions since awarding an appropriate role model and personal interest would require advance knowledge of the user's preferences. The current approach appeared more ecologically valid, e.g., giving your demographic information or your personal interests both could change the outcome of the experiment and is not typical information a game would have. Choice is also generally beneficial, e.g., [266, 301, 355, 472].

Quantitative and Qualitative Measures

Player Experience of Need Satisfaction

We use the 21-item Player Experience of Need Satisfaction (PENS) scale [474] that measures the following dimensions: Competence, Autonomy, Relatedness, Presence/Immersion, and Intuitive Controls. PENS is based on self-determination theory (SDT) [116]. PENS contends that the psychological “pull” of games are largely due to their ability to engender three needs—*competence* (seek to control outcomes and develop mastery [547]), *relatedness* (seek connections with others [30]), and *autonomy* (seek to be causal agents [94] while maintaining congruence with the self) [474]. PENS is considered a robust framework for assessing player experience [124, 458].

Computer Programming Self-Efficacy Scale

Self-efficacy represents the belief in one's ability to succeed, either in a particular situation, or at a particular task [24]. The Computer Programming Self-Efficacy Scale (CPSES) is a scale for measuring programming self-efficacy. It consists of a validated 32-item scale that measures the following dimensions: Independence and persistence, Complex programming tasks, Self-regulation, and Simple programming tasks [443].

Intrinsic Motivation Inventory

The Intrinsic Motivation Inventory (IMI) assesses intrinsic motivation using four dimensions: 1) Interest/Enjoyment, e.g., I enjoyed doing this activity very much, 2) Effort/Importance, e.g., I put a lot of effort into this, 3) Pressure/Tension, e.g., I felt very tense while doing this activity, 4) Value/Usefulness, e.g., I believe this activity could be of some value to me [367].

Player Inventory Scale

The Player Inventory Scale (PIS) measures avatar identification [526], which consists of three second-order factors: 1) Similarity identification, e.g., My character is similar to me, 2) Embodied identification, e.g., In the game, it is as if I become one with my character, 3) Wishful identification, e.g., I would like to be more like my character.

Map Quality Ratings

We collected both user and expert quality ratings of the final created game levels. Users were asked to rate their final game level on the dimensions of: “Aesthetic” (Is it visually appealing?), “Originality” (Is it creative?), “Fun” (Is it fun to play?), “Difficulty” (Is it difficult to play?), and “Overall” (Is it excellent overall?) on a scale of 1: *Strongly Disagree* to 7: *Strongly Agree*.

Expert ratings were given by 3 QA testers we hired. All QA testers had extensive games QA experience. The 3 QA testers first underwent supervised training in which they finished the game and created at minimum 3 maps in the editor. QA testers were then given 100 maps at random to establish baseline expectations. Next, QA testers were given another 25 maps at random to rate on the same dimensions as user ratings. Researchers verified the ratings and maps were rescored until there was consensus.

All 3 QA testers were blind to the experiment—the only information they received was a list of maps and links to each game level. They were debriefed on the purpose of their work after they completed all 2189 ratings. The 3 QA testers each spent an average of 109 hours (SD=8.5) over a 1-month period, at \$15 USD/hr.

Time Played

We directly measure motivation as operationalized by the amount of time spent playing the game and the editor.

Participants

2189 participants were recruited through Mechanical Turk. The data set consisted of 1001 male, and 1188 female participants. Participants self-identified their races/ethnicities as white (1681), black or African American (177), Chinese (41), Asian Indian (35), American Indian (23), Filipino (22), Korean (17), Vietnamese (13), Japanese (10) and other (170). Participants were between the ages of 18 and 71 ($M = 30.1$, $SD = 9.1$), and were all from the United States. Participants were reimbursed \$3.00 to participate in this experiment.

Design

A between-subjects design was used: badge condition was the between-subject factor. Participants were randomly assigned to a condition.

Protocol

Players first spent a minimum 4 minutes creating their Mii avatar. See Figure 5-37. The Mii creator was adapted from a freely available online Mii creator [104]. Options available included avatar/user name, gender, birthday, height, weight, favorite color (6), face shape (8), skin color (6), facial features (12), hair (72), hair color (8), eyebrows (24), eyebrow color (8), eyes (48), eye color (6), glasses (9), glasses color (6), nose (12), mouth (24), mouth color (3), mustache (4), beard (4), facial hair color (8), mole (2). There were also miscellaneous other options such as direction of hair part, and positioning, scale, and rotation of various facial elements. Avatar customization has previously been shown to increase, and produce a range of, avatar identification in users [23, 47, 334, 521].

Next, users completed the avatar identification scale. Before proceeding to the game, players were informed that they could exit the game *at any time* via a red button in the corner of the screen. Participants then played Mazzy. When participants were done playing (either by exiting early, or by finishing all 12 levels), participants completed the avatar identification

Continue in: 3:59



Figure 5-37: Mii avatar creator.

scale, the player experience of need satisfaction scale, the intrinsic motivation inventory, and the programming self-efficacy scale.

Next, users completed a tutorial which introduced them to the editor. The tutorial stepped users through all the interface elements and all editor functionalities. Each of the total 38 steps of the tutorial asked the user to perform an action before they could proceed, e.g., click a highlighted button to test the level. Each tutorial step had an additional help facility that provided additional troubleshooting information. After users completed the tutorial, they were required to spend at least 10 minutes creating a game level. After the 10 minutes passed, they could exit the editor via a red button in the corner of the screen, or continue using the editor until they wanted to quit. Users could repeat the tutorial at any time. After users quit the editor, they took a final screen capture of their level. This was done by positioning the map in the viewport and clicking a “Take Screenshot” button.

Users then completed the avatar identification scale, the player experience of need satisfaction scale, and the intrinsic motivation inventory. Users provided self-ratings on their

completed game levels, and filled out demographics.

Analysis

Data was extracted and imported into Statistical Package for Social Science (SPSS) version 22 for data analysis using multivariate analysis of variance (MANOVA). Separate MANOVAs are run for each separate set of items—*PENS*, *CPSES*, *IMI*, *PIS*; with the independent variable—*badge condition*. All the dependent variables are continuous variables. The independent variable badge condition (i.e., 0 = role model, 1 = personal interest, 2 = achievement, 3 = choice, 4 = choice with badges always visible, 5 = no badges) is a sexchotomous variable. To detect the significant differences between badge conditions, we utilized one-way MANOVA. These results are reported as significant when $p < 0.05$ (two-tailed). Prior to running our MANOVAs, we checked both assumption of homogeneity of variance and homogeneity of covariance by the test of Levene's Test of Equality of Error Variances and Box's Test of Equality of Covariance Matrices; and both assumptions were met by the data ($p > .05$ for Levene's, and $p > .001$ for Box's).

To measure the predictive capacity of avatar identification, we used linear hierarchical regression using similarity identification, embodied identification, and wishful identification as individual predictors. Since age and sex have been shown to affect need satisfaction, intrinsic motivation, and other avatar identification-related outcomes during game play [334, 456, 474], we entered age and sex in the first block of the regressions. We use avatar identification to predict game-related PENS, IMI, CPSES scores and editor-related PENS and IMI scores (using the avatar identification recorded pre-gameplay, and pre-editor, respectively). We then use avatar identification to predict play time and other time-related outcomes. Finally, we test if avatar identification can predict both self and expert ratings of final game level quality.

Results & Findings (Culminating Experiment)

RQ1: Do badges improve identification, need satisfaction, intrinsic motivation, and programming self-efficacy in the CS programming game and the subsequent game-making task?

Avatar Identification

Personal interest and role model badges promoted avatar identification in the game-making task. The MANOVA was not statistically significant across badge conditions in the game, $p > .05$. The MANOVA was statistically significant across badge conditions in the editor, $F(15, 6021) = 4.05$, $p < .0001$; Wilk's $\lambda = 0.973$, partial $\eta^2 = 0.01$. ANOVAs found the effect to be significant across all three dimensions of identification, $p < .0001$. Posthoc testing was done using Tukey HSD⁸. See Figure 5-38 and 5-39.

Similarity Identification:

- Personal Interest > No Badges (*editor*), $p < .05$
- Personal Interest > Achievement (*editor*), $p < .005$
- Personal Interest > Choice Always Visible (*editor*), $p < .001$
- Role Model > No Badges (*editor*), $p < .05$
- Role Model > Achievement (*editor*), $p < .005$
- Role Model > Choice Always Visible (*editor*), $p < .001$

Embodiment Identification:

- No Badge > Choice Always Visible (*editor*), $p < .05$
- Personal Interest > Achievement (*editor*), $p < .005$
- Personal Interest > Choice (*editor*), $p < .05$
- Personal Interest > Choice Always Vis. (*editor*), $p < .0001$
- Role Model > Achievement (*editor*), $p < .005$
- Role Model > Choice (*editor*), $p < .05$

⁸Note that we are calculating the *difference* in identification scores, i.e., post-editor minus pre-editor.

- Role Model > Choice Always Visible (*editor*), $p < .0001$

Wishful Identification:

- Role Model > Achievement (*editor*), $p < .05$
- Role Model > Choice (*editor*), $p < .01$
- Role Model > Choice Always Visible (*editor*), $p < .01$

Player Experience of Need Satisfaction

Achievement badges promoted player experience in the CS programming game. Role model badges promoted player experience in both the CS programming game and the game-making task. The MANOVA was statistically significant across badge conditions in the game and the editor, $p < .05$. ANOVAs found that the effect was significant across all five dimensions in both the game and the editor, $p < .05$. Posthoc testing using Tukey HSD found that for **competence**: Role Model > Personal Interest (*game*), $p < .05$, Role Model > Choice (*editor*), $p < .005$, Role Model > Choice Always Visible (*editor*), $p < .01$. For **autonomy**: Achievement > Personal Interest (*game*), $p < .05$, Role Model > Personal Interest (*game*), $p < .05$, Role Model > Choice (*editor*), $p < .05$, Role Model > Choice Always Visible (*editor*), $p < .05$. For **relatedness**: Achievement > Personal Interest (*game*), $p < .05$, Role Model > Personal Interest (*game*), $p < .05$. For **immersion**: Achievement > Personal Interest (*game*), $p < .05$, Role Model > Personal Interest (*game*), $p < .05$, Role Model > Choice Always Visible (*editor*), $p < .05$. For **intuitive control**: Role Model > Choice (*game*), $p < .05$, Role Model > Choice (*editor*), $p < .05$. See Figure 5-38 and 5-39.

Intrinsic Motivation Inventory

Achievement badges promoted intrinsic motivation in the CS programming game. Role model badges promoted intrinsic motivation in both the CS programming game and the game-making task. The MANOVA was statistically significant across badge conditions in the game and the editor, $p < .05$. ANOVAs found that the effect was significant for the dimensions of enjoyment (*game, editor*), $p < .05$, usefulness (*editor*), $p < .05$. Posthoc testing using Tukey HSD found that for **enjoyment**: Achievement > No Badges (*game*),

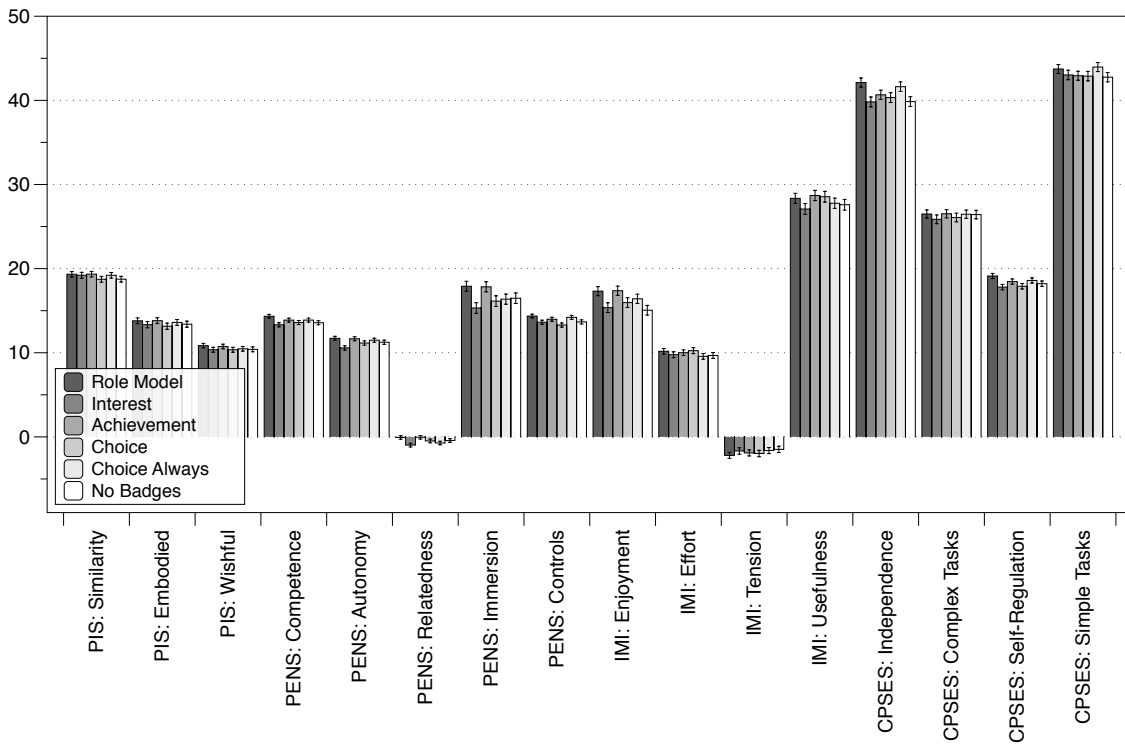


Figure 5-38: Measures, Post-Game. Error bars are standard error of the mean (SEM).

$p < .05$, Role Model > No Badges (*game*), $p < .05$, Role Model > Choice Always Visible (*editor*), $p < .01$. For **usefulness**: Role Model > Choice Always Visible (*editor*), $p < .05$. See Figure 5-38 and 5-39.

Computer Programming Self-Efficacy Scale

Role model badges promoted programming self-efficacy in the CS programming game.

The MANOVA was statistically significant across badge conditions, $p < .0001$. ANOVAs found that the effect was significant for the dimensions of independence (*game*), $p < .05$, and self-regulation (*game*), $p < .05$. Posthoc testing using Tukey HSD found that for **independence**: Role Model > Personal Interest (*game*), $p < .05$. For **self-regulation**: Role Model > Personal Interest (*game*), $p < .05$. See Figure 5-38.

Time Played

ANOVAs found no significant effect of badge condition on **game time**: $M=1497.06$, $SD=1952.58$, $F(5, 2183)=0.64$, $p=0.67$; and **editor time**: $M=741.02$, $SD=723.44$, $F(5,$

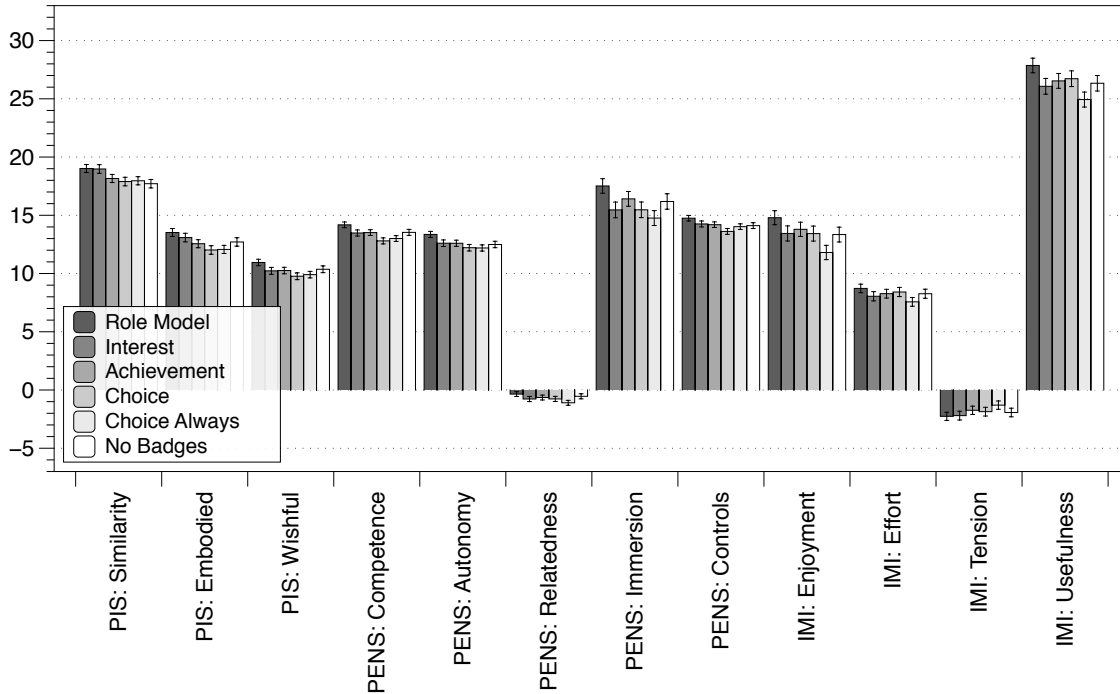


Figure 5-39: Measures, Post-Editor. Error bars show SEM.

2183)=0.88, $p=0.49$.

RQ2: Does avatar identification improve need satisfaction, intrinsic motivation, and programming self-efficacy in the CS programming game and the subsequent game-making task?

From the hierarchical regression in Table 5.18, avatar identification significantly improves need satisfaction, intrinsic motivation, and programming self-efficacy in both the game and the editor. On average, significant R^2 values explain 7.4% of variance.

RQ3: Does avatar identification translate into higher motivated behavior (time spent, etc.)?

From the hierarchical regression in Table 5.18, embodied identification leads to higher game time and similarity identification leads to higher editor time. All dimensions of avatar identification lead to more time playtesting in editor, and more time spent taking the final screenshot. On average, significant R^2 values explain 1.8% of variance.

	Similarity Identification					Embodied Identification					Wishful Identification				
	β	R^2	$R^2(c)$	$F(c)$	$p(c)$	β	R^2	$R^2(c)$	$F(c)$	$p(c)$	β	R^2	$R^2(c)$	$F(c)$	$p(c)$
Player Experience of Need Satisfaction (PENS)															
Competence (<i>game</i>)	0.13	0.024	0.016	35.4	0.00	0.00	0.009	0.000	0.00	0.96	0.13	0.030	0.022	49.1	0.00
Competence (<i>editor</i>)	0.25	0.069	0.061	142	0.00	0.23	0.060	0.053	123	0.00	0.17	0.037	0.028	64.5	0.00
Autonomy (<i>game</i>)	0.19	0.034	0.034	76.3	0.00	-0.01	0.000	0.000	0.05	0.82	0.24	0.059	0.059	136	0.00
Autonomy (<i>editor</i>)	0.28	0.077	0.075	178	0.00	0.32	0.104	0.102	248	0.00	0.25	0.065	0.063	148	0.00
Relatedness (<i>game</i>)	0.18	0.034	0.030	68.2	0.00	-0.02	0.004	0.000	0.81	0.37	0.33	0.112	0.109	267	0.00
Relatedness (<i>editor</i>)	0.21	0.050	0.044	101	0.00	0.48	0.234	0.228	650	0.00	0.41	0.177	0.171	453	0.00
Immersion (<i>game</i>)	0.24	0.060	0.059	136	0.00	-0.02	0.002	0.000	0.84	0.36	0.40	0.163	0.162	422	0.00
Immersion (<i>editor</i>)	0.33	0.108	0.108	264	0.00	0.63	0.394	0.394	1418	0.00	0.51	0.257	0.257	755	0.00
Intuitive Control (<i>game</i>)	0.10	0.021	0.009	20.9	0.00	-0.01	0.012	0.000	0.09	0.76	0.11	0.024	0.012	27.6	0.00
Intuitive Control (<i>editor</i>)	0.22	0.056	0.048	111	0.00	0.20	0.049	0.041	93.7	0.00	0.13	0.026	0.018	39.9	0.00
Intrinsic Motivation Inventory (IMI)															
Enjoyment (<i>game</i>)	0.19	0.041	0.035	79.8	0.00	0.00	0.006	0.000	0.01	0.93	0.17	0.035	0.029	65.9	0.00
Enjoyment (<i>editor</i>)	0.24	0.059	0.058	135	0.00	0.29	0.084	0.083	198	0.00	0.22	0.051	0.050	115	0.00
Effort (<i>game</i>)	0.20	0.080	0.038	91.2	0.00	0.02	0.042	0.001	1.17	0.28	0.12	0.056	0.015	33.9	0.00
Effort (<i>editor</i>)	0.23	0.088	0.051	122	0.00	0.21	0.080	0.043	103	0.00	0.16	0.061	0.026	60.1	0.00
Tension (<i>game</i>)	0.02	0.004	0.000	1.03	0.31	-0.01	0.003	0.000	0.17	0.68	0.07	0.008	0.005	10.3	0.00
Tension (<i>editor</i>)	-0.01	0.005	0.000	0.07	0.79	0.12	0.019	0.014	31.4	0.00	0.12	0.019	0.014	32.1	0.00
Usefulness (<i>game</i>)	0.21	0.043	0.042	95.8	0.00	0.00	0.001	0.000	0.00	0.99	0.26	0.065	0.065	151	0.00
Usefulness (<i>editor</i>)	0.31	0.096	0.094	228	0.00	0.38	0.149	0.147	376	0.00	0.32	0.106	0.104	254	0.00
Computer Programming Self-Efficacy Scale (CPSES)															
Independence (<i>game</i>)	0.06	0.026	0.004	8.54	0.00	0.01	0.023	0.000	0.31	0.57	0.02	0.023	0.000	0.60	0.44
Complex Tasks (<i>game</i>)	0.08	0.035	0.005	12.4	0.00	0.01	0.030	0.000	0.07	0.80	0.08	0.035	0.006	12.8	0.00
Self-Regulation (<i>game</i>)	0.09	0.018	0.009	19.4	0.00	-0.01	0.011	0.000	0.43	0.51	0.07	0.016	0.005	11.2	0.00
Simple Tasks (<i>game</i>)	0.08	0.020	0.006	12.4	0.00	0.01	0.014	0.000	0.18	0.67	0.02	0.015	0.000	0.92	0.34
Behavior															
Time Played (<i>game</i>)	0.03	0.017	0.001	1.45	0.23	0.04	0.019	0.002	4.20	0.04	0.02	0.017	0.000	0.56	0.46
Time Played (<i>editor</i>)	0.04	0.002	0.002	4.13	0.04	0.02	0.001	0.000	0.71	0.40	-0.02	0.001	0.001	1.10	0.30
Time Testing (<i>editor</i>)	0.02	0.026	0.000	0.87	0.35	0.07	0.030	0.005	11.0	0.00	0.06	0.029	0.004	8.79	0.00
Time Taking Screenshot (<i>editor</i>)	0.08	0.017	0.006	12.3	0.00	0.06	0.015	0.004	8.24	0.00	0.07	0.016	0.005	10.1	0.00
Game Map Ratings															
Overall Quality (<i>self-rated</i>)	0.20	0.059	0.038	87.2	0.00	0.23	0.074	0.052	124	0.00	0.20	0.063	0.041	95.2	0.00
Aesthetic (<i>expert-rated</i>)	0.06	0.021	0.004	7.86	0.00	0.07	0.021	0.004	9.45	0.00	-0.04	0.019	0.002	3.62	0.06
Originality (<i>expert-rated</i>)	0.05	0.021	0.002	5.04	0.03	0.05	0.022	0.003	6.44	0.01	-0.06	0.023	0.004	9.14	0.00
Fun (<i>expert-rated</i>)	0.04	0.024	0.002	3.98	0.05	0.06	0.025	0.003	6.81	0.01	-0.06	0.026	0.004	8.86	0.00
Difficulty (<i>expert-rated</i>)	0.05	0.037	0.003	6.09	0.01	0.03	0.035	0.001	1.98	0.16	-0.06	0.037	0.003	7.41	0.01
Overall Quality (<i>expert-rated</i>)	0.06	0.027	0.003	7.01	0.01	0.06	0.028	0.004	8.30	0.00	-0.06	0.027	0.003	7.70	0.01

Table 5.18: Regression properties β , R^2 , change in R^2 , F , and p from adding identification. Change statistics are marked (c). Significant results are bold.

RQ4: Does avatar identification improve created game levels?

Intraclass correlation across the three raters on overall quality was ICC=0.83 (two-way random, average measures), indicating high agreement. From the hierarchical regression in Table 5.18, all three dimensions of avatar identification lead to higher self-perceived quality. Similarity identification and embodied identification lead to increases in actual game level quality. However, wishful identification leads to a decrease in actual game level quality. On average, significant R^2 values explain 3.3% of variance. Average quality as rated by experts was M=3.54, SD=1.15. See Figure 5-40.

Experiment-Specific Discussion (Culminating Experiment)

We found that each of our research questions could be answered in the affirmative. Badges promoted avatar identification (interest, role model), player experience (achievement, role model), intrinsic motivation (achievement, role model), and programming self-efficacy (role

model) during both the game and the editor. Role model badges were particularly effective during the game making task.

Our results have implications for both play and making. We find that role model badges improve virtually all facets of experience (player experience, intrinsic motivation, self-efficacy) relative to other badge types. This effectiveness was particularly relevant during the game making task. Achievement badges were found to be effective in the game, which would corroborate previous work, e.g., [12, 121, 328, 390]. Personal interest badges were found to only improve avatar identification during the game making task. Both choice conditions did not appear to be effective—possibly indicative of too much choice, or that the choice between badge types was simply not a meaningful one [150, 164, 292, 462].

Badges appeared to differ in their effectiveness based on the game or the editor. Therefore, it's likely that task context is a moderator—e.g., achievement badges earned during the game are less effective in the editor, whereas role model badges may generalize across the two. We also note that for all conditions except the Choice Always Visible condition, badges were only briefly visible after having completed a level in game. It's possible that having badges visible at all times—as during the editor—would further reinforce badge effects.

Additionally, we found that avatar identification positively affects all measures (player experience, intrinsic motivation, programming self-efficacy, overall time) in both play and making. Furthermore, avatar identification leads to higher quality completed maps. Therefore, both badges and avatar identification affect a variety of play and making related outcomes.

One caveat, however, is that wishful identification was actually negatively correlated to map quality. Wishful identification—or wanting to be like a fictional or media character [157, 237, 238, 344]—is correlated with lower psychological well-being [44, 234, 387]. However, wishful identification may be beneficial for self-esteem [44, 140]. This two-sided nature of wishful identification was expressed here as a universally positive effect on outcomes, *except* on actual game level quality. More work on wishful identification is needed to precisely characterize why this was the case.

Our R^2 values range between small (0.01), medium (0.09), and large (0.25) effect sizes. Avatar identification was particularly predictive of player experience, e.g., similarity identification (10.8%), embodied identification (39.4%), and wishful identification (25.7%) were all highly predictive of immersion during game making. Our mean significant R^2 value is 5.9% which we've demonstrated at a scale of $N=2189$ across many different outcomes, suggesting that avatar identification is an important component to our play and making experience.

Badges Applications

Our results suggest role model badges are effective—similarly to role model avatars [277]—yet badges in contrast may have more general application. We might imagine scientific games that leverage the crowd (e.g., FoldIt [297]), MOOCs, digital learning platforms, etc. as being possible beneficiaries of these badges.

Literature suggests that role models are useful outside of academic contexts (e.g., [96])—as long as they are relevant [341]. Therefore, other domains such as business (Steve Jobs, etc.), politics (Barack Obama, etc.), health (Oprah Winfrey, etc.), may also benefit from these badges—a game, an educational platform, a gamified app, etc.—so long as the role models meet the criteria of perceived competence, similarity, and success.

Avatar Applications

Our results suggest that avatar identification can improve time on task, and positively impact player experience, intrinsic motivation, and self-efficacy. Applications range from enhancing motivation in crowdsourced tasks, to serious contexts such as behavior change, education, etc. We extend previous work [47] by showing that avatar identification can impact the quality of created levels. It remains an open and interesting question as to whether the production of other artifacts can also be similarly affected through identification with an on-screen representation—e.g., writing an essay, programming an application, designing a graphic, etc.

With increasing emphasis on making as a pedagogical method, there is ongoing concern about the quality of produced artifacts [118, 273]. Critics of Mario Maker, for instance,

have condemned the majority of user created levels as being impossible, gimmicky, and “bafflingly opaque, frenzied contraptions that rarely seem to have a purpose” [118, 230, 516]. Here, we have made a first step towards understanding badges and avatar identification in relation to creation.

Limitations

Our study consisted of a period of time on the order of hours. However, the interaction between a user and a game often extends into long-term (e.g., in World of Warcraft [339]). Therefore, a longitudinal study could elucidate how our results are moderated by longer term use.

We also took one specific approach to studying badges. For example, we decided that making the badges black and white (to control for color confounds [190, 374, 376]) was a necessary price to pay. However, future studies could introduce color in a controlled way to further understand how color can moderate our findings. Another example was in how we implemented role model badges. We were cautious to ensure that our badge creation process yielded: 1) role models perceived as competent [358], 2) role models perceived as ingroup [340, 357], and 3) role models perceived as successful [78]—the lattermost was additionally reinforced with in-game text. However, how much deviation from these criteria that can still result in effective role model badges remains to be explored.

Summary

In this study, we have looked at how badges and avatar identification impact both play and making in an educational game. We found that certain badges could promote avatar identification (personal interest, role model), player experience (achievement, role model), intrinsic motivation (achievement, role model), and programming self-efficacy (role model) during both the game and the editor.

Avatar identification promoted player experience, intrinsic motivation, programming self-efficacy, and the total time spent playing and making. Avatar identification also promoted other meaningful in-editor activity, such as playtesting time, etc. and led to significantly



Figure 5-40: Example maps rated overall 2 (left), 4 (center), and 6 (right).

higher overall quality of the completed game levels (as rated by 3 independent externally trained QA testers). Here, we've conducted a first study (N=2189) on alternative badge types, and a first study of badges and avatar identification in a making context. These findings contribute to both the literature on badges and avatars.

5.5 Chapter References

This chapter, in part, contains material that is a reprint of published papers.

Experiment [Shape vs. Likeness #1/#2](#) is described in the paper *Toward Avatar Models to Enhance Performance and Engagement in Educational Games* from the 2015 Computational Intelligence in Games (CIG) conference [282].

Experiment [Shape vs. RoleModel](#) is described in the paper *Exploring the Use of Role Model Avatars in Educational Games* from the 2015 Artificial Intelligence in Interactive Digital Entertainment (AIIDE) Experimental AI in Games workshop [279].

Experiment [Shape vs. Scientist vs. Athlete](#) is described in the paper *Exploring the Impact of Role Model Avatars on Game Experience in Educational Games* from the 2015 ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHI PLAY) conference [281]. Interested readers can refer to *Toward Understanding the Impacts of Role Model Avatars on Engagement in Computer Science Learning* from the 2016 American Educational Research Association (AERA) conference [285] for an extended analysis on the full (N=1067) dataset.

Experiment [Successful Likeness](#) is described in the paper *Exploring the Effects of Dynamic Avatar on Performance and Engagement in Educational Games* from the 2016 Games+Learning+Society (GLS) conference [286].

Experiment [Red vs. Blue](#) is described in the paper *Exploring the Impact of Avatar Color on Game Experience in Educational Games* from the 2016 Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA) [287].

Experiment [Feedback Positive vs. Negative vs. Neutral vs. Nothing](#) is described in the paper *Exploring the Effects of Encouragement in Educational Games* from the 2016 Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA) [288].

Experiment [Game Theme Basic vs. Circuit vs. RPG vs. Choice](#) is described in the paper *Toward Understanding the Impact of Visual Themes and Embellishment on Performance, Engagement, and Self-Efficacy in Educational Games* from the 2017 American Educational Research Association (AERA) conference [[289](#)].

Experiment [Badge Type Comparison](#) is described in the paper *The Effects of Badges and Avatar Identification on Play and Making in Educational Games* from the 2018 Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems (CHI) [[290](#)].

Chapter 6

Conclusion

6.1 Main Findings

I begin with a re-summarization of our findings:

Avatar-Based Outcomes:

- **Simple avatars often outperform complex avatars** [286].
- **Scientist role model avatars are extremely effective** [277, 279, 285].
- **Successful likeness avatars can likely outperform any existing avatar types** [286].
- **Red avatars cause significant decreases in engagement and avatar affect compared to blue avatars** [287].
- **Badges and avatar identification promote positive outcomes** [290].

Other Outcomes:

- **Positive and neutral encouragement text displayed at regular intervals (e.g., “Keep it up!”), significantly increases engagement as compared to no text or negative encouragement text** [288].
- **More embellished game backgrounds cause players to have significantly decreased game performance and significantly decreased programming self-efficacy**

but significantly increased engagement [289].

6.2 Domain Applications

The findings from this thesis can most directly be applied to educational contexts, e.g., adaptive learning, educational games, MOOCs, etc. However, the implications may be much broader. We can imagine virtual representations playing a role in *crowdsourcing*, e.g., scientific games that leverage the crowd such as FoldIt [297], *behavior change*, e.g., apps for smoking cessation [6], *virtual reality*, e.g., VR re-creation of the Titanic [1], etc.—all of which can benefit from identification with an avatar, and avatar types more conducive to performance, engagement, etc. We could imagine the experience of playing Marie Curie in a lab. We could imagine avatars that shift between likeness, famous person, and abstract depending on context. We could imagine simulations that change the color of the avatar, perhaps through lighting, to induce a particular state. We could imagine digital experiences that adapt degree of embellishment depending on the goals of the user. We could imagine that writing an essay, programming an application, designing a graphic, etc. might be improved through identification with an avatar. We could imagine that role model badges might be applied as widely as achievement badges. All of these could serve to reinforce different aspects of user and interface.

6.3 Limitations

Despite that our findings can be applied widely, there is much to be done. Generally, our experiments have been limited to be on the order of hours. Long-term controlled experiments would further elucidate the moderating effects of time. Work has shown that time increases congruence/identification with avatars [140, 445, 521, 560, 562]—therefore time *may* positively reinforce pre-existing impacts.

In exploring avatar types, we have only tapped a small subspace of potential avatars. The

possible avatars are uncountable—e.g., animals, furniture objects, fantastical dragons, institutions like colleges, corporations, oceans, etc. While exploring *every* representation is an unnecessary expenditure of resources—with sufficient data, modeling effects will be possible—due diligence to the infinite space of representation is important.

This thesis has presented results from crowdsourced studies of more than 10,000 participants. We have studied avatars, embellishment, badges, etc., and our evidence demonstrates meaningful effects on educational outcomes. I conclude with potential future directions.

6.4 Future Work

In our experiment on the successful likeness, we demonstrated that the effectiveness of an avatar can vary by the current situation within a single game—specifically, we showed that an avatar that represented the user during success, but that represented an object during failure, was more optimal. Therefore, our belief is that machine learning algorithms that use context—the individual’s demographics, the context, the individual’s preferences, etc.—have great potential. Imagine an avatar that is constantly shifting in a subtle way depending on the individual’s emotional and cognitive state in order to both increase their mood and to elevate cognition for the current task. With machine learning, it becomes possible to adapt the avatar to optimize for variables of interest, e.g., self-efficacy, etc. with enough historical data.

We have demonstrated that avatars’ impacts are a complex interplay of variables. We have drawn from literature as varied as color perception, stereotypes, and character identification. However, further elucidating an all-encompassing model, e.g., exploring physiological differences between avatar types, such as fMRI data, would be one approach to mitigate the number of representations in an infinite space. Eventually, mathematical models can formally represent avatars’ impacts as a function of color, user demographics, task context, etc. and predict short and long term impacts.

That being said, not all effects will be quantifiable. Degree of identification with a subject matter for instance, may well be affected through virtual representation, but may not be reflected in existing assessments. It cannot be stated highly enough that virtual identity is highly situated—its impactfulness varies by individual and environmental context. As such, there is also a need for qualitative, humanistic, and artistic methods.

6.5 Closing Reflections

We have contributed a systematic series of studies that expands our knowledge in the domain of avatars and learning. We have developed design principles for makers of educational environments, and digital contexts more generally. This domain will become increasingly important as our world becomes more digital. What will the classrooms, games, and work environments of the future—a future filled with ubiquitous virtual and augmented reality—look like? Our work suggests that whatever form these environments take on, that the virtual representation we (and others) use have significant implications.

We contend that most systems can benefit from the results discussed in this thesis. Software without existing virtual representations could be augmented with, for instance, a helper agent. User profiles, emotes, and chat icons are some of the other methods being used to express the self. What form these representations take on are set to shape us in new ways.

“Avatars” are all around us. They are our social media accounts, our virtual bodies in the Oculus Rift, our online Amazon accounts. These representations are as malleable as our physical identities. We are always stepping in out of different avatars. Both the ubiquity and impacts of avatars gives us strong cause to continue our pursuit of knowledge in this domain. These studies are a modest step into a vast and important domain in which we seek to understand the impacts of the signs we use to represent ourselves on our learning-in-the-world.

Appendix A

Mazzy Version Listing

Version 0 – Same as #1 without background music.

Version 1 – Described in Section 3.1.6.

http://groups.csail.mit.edu/icelab/mazzy_versions/v1



System requirements: Unity Plugin (e.g., Safari, Internet Explorer, etc.).

Gameplay video: <http://youtu.be/j0TI4MH2rsY>

Version 2 – Prototype version of #4.

http://groups.csail.mit.edu/icelab/mazzy_versions/v2



System requirements: Unity Plugin (e.g., Safari, Internet Explorer, etc.).

Note: Version used in experiments enforced command limits for each level.

Version 3 – Same as #2 + revamped UI.

Note: Also added a congratulatory message with player's avatar centered in middle of screen after each level completion. Players could also no longer redo levels after completing them (as they could in #2).

Version 4 – Described in Section 3.1.

http://groups.csail.mit.edu/icelab/mazzy_versions/v4



System requirements: WebGL (e.g., Chrome, Firefox, etc.).

Note: Between-level surveys were removed. A “stars” screen that awarded stars based on command count was removed.

Gameplay video: <http://youtu.be/n2rR1CtVal8>

Table A.1: Mazzy Versions.

Appendix B

Calculations

Game Version #0, #1:

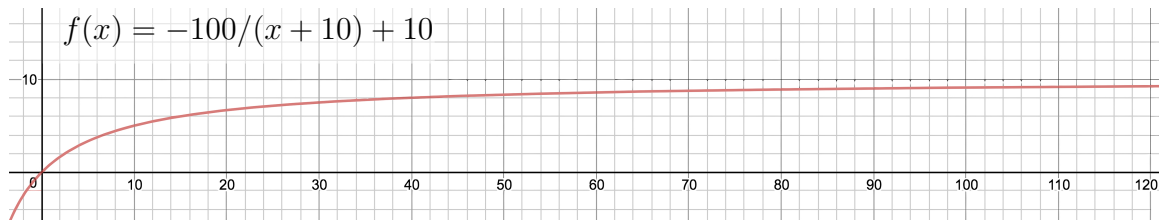
$$Performance = 30\frac{1}{3} * (LevelsCompleted) + 1 * (BonusCapsules)$$

$$Engagement = 0.9 * (ReportedMean) + 0.1 * (PlaytimeFunction)$$

Game Version #2, #3, #4:

$$Performance = 8\frac{1}{3} * (LevelsCompleted)$$

$$Engagement = 0.9 * (ReportedMean) + 0.1 * (PlaytimeFunction)$$

All Game Versions:

$$PlaytimeFunction = f(Percentile) + if(Percentile > 5)\{100/110\}$$

Percentile is the participant's playtime percentile with respect to *all other participants in the same game version*. This function intentionally penalizes participants that played very little (comparatively). Note that the limit of $f(x)$ as x goes to infinity is 10, but since *Percentile's* max is 100, we include the second term to make the range of the result $\{0, 10\}$.

Appendix C

Protocol Versions

#1: Choose between eight shapes¹

#2: Create a Mii

#3: None²

#4: Find face photo

#5: Choose between eight scientists (4 CEOs of tech companies: Bill Gates, Steve Jobs, Larry Page, Mark Zuckerberg, 4 Google³ scientists: Stephen Hawking, Galileo, Marie Curie, Einstein).

#6: Choose between 15 scientists (Google)

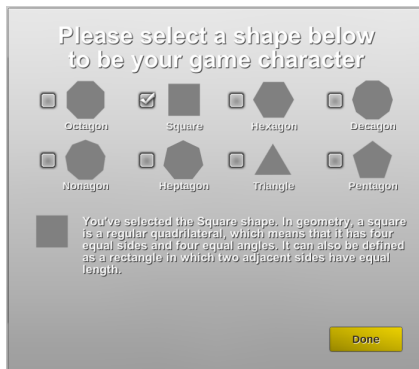
#7: Google a shape image

#8: Google a role model image

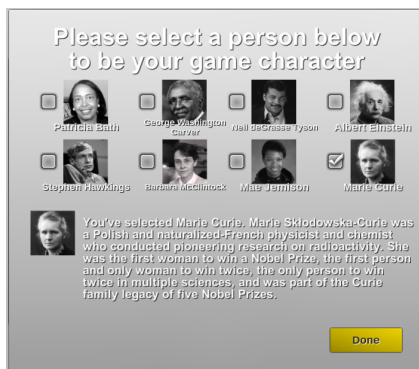
#9: Choose 3 affirmations⁴

#10: Choose 3 non-affirmations

#11: Choose between eight shapes



#12: Choose between eight scientists



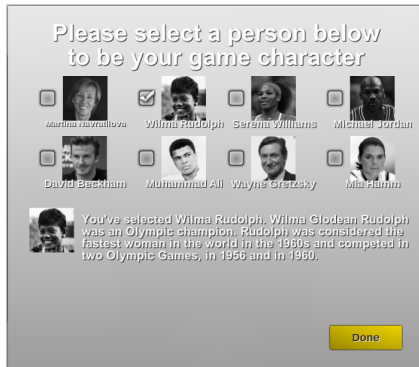
¹Pink shapes.

²In other words, there was no avatar selection/creation/customization whatsoever, the player jumps into the game and has whatever avatar we have assigned in that case.

³Google's top results

⁴These affirmations then appeared *on* the avatar itself as text (e.g., athletic ability)

#13: Choose between eight athletes



#14: Create a Mii (#2) and choose between eight shapes (#11)

#15: Customize avatar in built-in creator, plus *phantoms* of other players also appear intermittently during the game

#16: Customize avatar in built-in creator

#17: Customize avatar in built-in creator, plus *positive feedback* as text

#18: Customize avatar in built-in creator, plus *negative feedback* as text

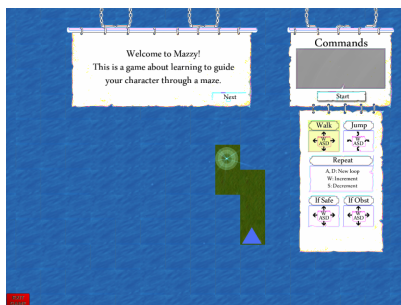
#19: Customize avatar in built-in creator, plus *neutral feedback* as text

#20: Customize avatar in built-in creator, afterwards losing at a mini-game

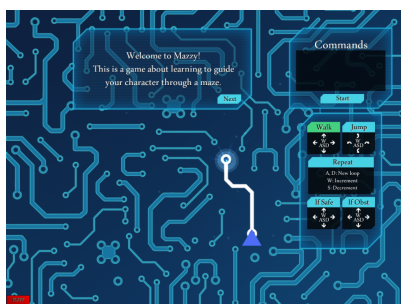
#21: Customize avatar in built-in creator, afterwards *nearly winning* the mini-game

#22: Customize avatar in built-in creator, afterwards winning the mini-game

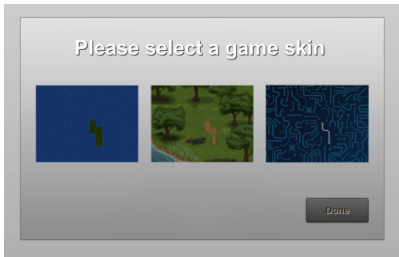
#23: None; play the game with a basic-looking skin



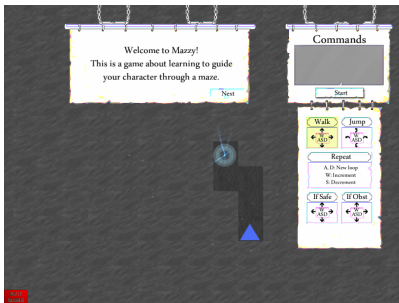
#24: None; play the game with a circuitboard skin



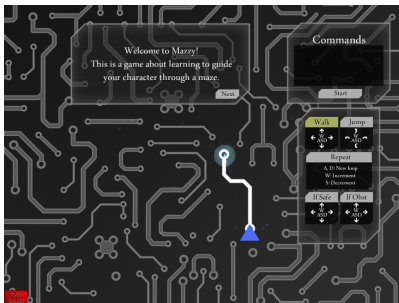
#25: None; play the game with choice of 3 skins



#26: None; play the game with a black/white basic-looking skin



#27: None; play the game with a black/white circuitboard skin



#28: None; play the game with a black/white RPG-like skin



#29: None; play the game with choice of 3 black/white skins

Appendix D

Additional Tables

Factor 1: Independence and persistence (alpha = 0.84)

1. Complete a program if I had no help at all.
2. Complete a program once the tutorial helped me get started.
3. Complete a program if someone showed me how to solve the problem first.

Factor 2: Complex programming tasks (alpha = 0.85)

1. Write a program for an extremely difficult problem.
2. Organize my program in a clean way.
3. Mentally trace through the execution of a long, complex, program given to me.

Factor 3: Self-regulation (alpha = 0.85)

1. Come up with a suitable strategy for a given problem in a short time.
2. Manage my time efficiently if I had a pressing deadline on a problem.
3. Find a way to concentrate on my program, even when there were many distractions around me.

Factor 4: Simple programming tasks (alpha = 0.86)

1. Write logically correct blocks of code.
2. Write a program for a simple problem.
3. Write a program for a moderately difficult problem.

Table D.1: CPSES

Effect		Value	Hypothesis			Partial Eta	
			<i>F</i>	<i>df</i>	Error <i>df</i>	Sig.	Squared
Intercept	Pillai's Trace	.871	2615.055 ^a	3.000	1165.000	.000	.871
	Wilks' Lambda	.129	2615.055 ^a	3.000	1165.000	.000	.871
	Hotelling's Trace	6.734	2615.055 ^a	3.000	1165.000	.000	.871
	Roy's Largest	6.734	2615.055 ^a	3.000	1165.000	.000	.871
	Root						
NumericCondition	Pillai's Trace	.092	12.344	9.000	3501.000	.000	.031
	Wilks' Lambda	.908	12.777	9.000	2835.454	.000	.032
	Hotelling's Trace	.102	13.128	9.000	3491.000	.000	.033
	Roy's Largest	.101	39.277 ^b	3.000	1167.000	.000	.092
	Root						

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept + NumericCondition

Table D.2: Performance—MANOVA Multivariate F-tests

Dependent Variable	Condition	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Levels Completed	Generic	8.187	.178	7.837	8.536
	Choice	8.010	.182	7.654	8.367
	Fantasy	7.803	.177	7.456	8.150
	Circuit	7.470	.185	7.107	7.832
Hints Requested	Generic	2.003	.183	1.644	2.362
	Choice	2.451	.187	2.085	2.818
	Fantasy	2.510	.182	2.153	2.867
	Circuit	3.172	.190	2.800	3.544
Attempts	Generic	24.397	1.730	21.002	27.791
	Choice	32.174	1.766	28.709	35.638
	Fantasy	32.421	1.719	29.049	35.793
	Circuit	41.086	1.794	37.566	44.606

Table D.3: Performance—Descriptive

Dependent Variable	Conditions	p-value
Levels Completed	Generic > Circuit	p < .005
Levels Completed	Choice > Circuit	p < .05
Hints Requested	Generic < Fantasy	p < .05
Hints Requested	Generic < Circuit	p < .001
Hints Requested	Choice < Circuit	p < .01
Hints Requested	Fantasy < Circuit	p < .05
Attempts	Generic < Choice	p < .001
Attempts	Generic < Fantasy	p < .001
Attempts	Generic < Circuit	p < .001
Attempts	Choice < Circuit	p < .001
Attempts	Fantasy < Circuit	p < .001
Attempts	Fantasy < Circuit	p < .001

Post hoc comparisons (LSD) revealed that participants in the generic condition completed more levels than participants in the circuit condition, $p < .005$. Participants in the choice condition also completed more levels than participants in the circuit condition, $p < .05$. Participants in the generic condition used less hints than participants in either the fantasy, $p < .05$, or circuit, $p < .001$, conditions. Participants in the choice condition used less hints than participants in the circuit condition, $p < .01$. Participants in the fantasy condition used less hints than participants in the circuit condition, $p < .05$. Participants in the generic condition used less attempts than participants in the choice, fantasy, or circuit conditions, $p < .001$. Participants in the choice condition used less attempts than participants in the circuit condition, $p < .001$. Participants in the fantasy condition used less attempts than participants in the circuit condition, $p < .001$.

Table D.4: Performance—Posthocs

Effect		Hypothesis				Partial Eta	
		Value	<i>F</i>	<i>df</i>	Error <i>df</i>	Sig.	Squared
Intercept	Pillai's Trace	.934	1358.838 ^a	12.000	1156.000	.000	.934
	Wilks' Lambda	.066	1358.838 ^a	12.000	1156.000	.000	.934
	Hotelling's Trace	14.106	1358.838 ^a	12.000	1156.000	.000	.934
	Roy's Largest Root	14.106	1358.838 ^a	12.000	1156.000	.000	.934
	Root						
NumericCondition	Pillai's Trace	.045	1.472	36.000	3474.000	.035	.015
	Wilks' Lambda	.956	1.471	36.000	3416.259	.035	.015
	Hotelling's Trace	.046	1.470	36.000	3464.000	.035	.015
	Roy's Largest Root	.019	1.842 ^b	12.000	1158.000	.038	.019
	Root						

a. Exact statistic

b. The statistic is an upper bound on *F* that yields a lower bound on the significance level.

c. Design: Intercept + NumericCondition

Table D.5: Self-Efficacy—MANOVA Multivariate *F*-tests

Dependent Variable	Condition	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Answer.EfficacyQ1	Generic	4.057	.112	3.836	4.277
	Choice	4.066	.115	3.841	4.291
	Fantasy	3.954	.112	3.735	4.173
Answer. EfficacyQ2	Circuit	3.663	.117	3.434	3.892
	Generic	5.213	.098	5.022	5.405
	Choice	5.267	.100	5.072	5.463
	Fantasy	5.273	.097	5.083	5.463
Answer. EfficacyQ3	Circuit	5.000	.101	4.801	5.199
	Generic	5.267	.103	5.064	5.469
	Choice	5.392	.105	5.186	5.599
	Fantasy	5.431	.103	5.230	5.632
Answer. EfficacyQ4	Circuit	5.108	.107	4.898	5.318
	Generic	3.300	.108	3.089	3.511
	Choice	3.042	.110	2.826	3.257
	Fantasy	3.076	.107	2.866	3.286
Answer. EfficacyQ5	Circuit	2.824	.112	2.605	3.044
	Generic	4.220	.110	4.005	4.435
	Choice	4.090	.112	3.871	4.310
	Fantasy	3.918	.109	3.704	4.131
Answer. EfficacyQ6	Circuit	3.932	.114	3.709	4.155
	Generic	4.303	.106	4.096	4.511
	Choice	4.132	.108	3.920	4.344
	Fantasy	3.980	.105	3.774	4.186
Answer. EfficacyQ7	Circuit	3.932	.110	3.717	4.147
	Generic	3.973	.099	3.779	4.168
	Choice	3.799	.101	3.600	3.997
	Fantasy	3.822	.098	3.629	4.015
Answer. EfficacyQ8	Circuit	3.559	.103	3.358	3.760
	Generic	4.060	.099	3.866	4.254
	Choice	3.851	.101	3.652	4.049
	Fantasy	3.789	.098	3.596	3.983
Answer. EfficacyQ9	Circuit	3.563	.103	3.361	3.764
	Generic	4.433	.107	4.224	4.643
	Choice	4.295	.109	4.082	4.509
	Fantasy	4.263	.106	4.055	4.471
Answer. EfficacyQ10	Circuit	4.222	.111	4.005	4.439
	Generic	4.790	.104	4.587	4.993
	Choice	4.462	.106	4.254	4.669
	Fantasy	4.599	.103	4.397	4.801
Answer. EfficacyQ11	Circuit	4.419	.107	4.209	4.630
	Generic	5.677	.109	5.463	5.890
	Choice	5.552	.111	5.334	5.770
	Fantasy	5.589	.108	5.377	5.801
Answer. EfficacyQ12	Circuit	5.233	.113	5.012	5.454
	Generic	4.707	.108	4.495	4.918
	Choice	4.469	.110	4.253	4.685
	Fantasy	4.480	.107	4.270	4.691
	Circuit	4.172	.112	3.953	4.392

Table D.6: Self-Efficacy—Descriptives

Dependent Variable	Conditions	p-value
Complete a program if I had no help at all	Generic > Circuit	p < .05
Complete a program if I had no help at all	Choice > Circuit	p < .05
Write a program for an extremely difficult problem	Generic > Circuit	p < .05
Come up with a suitable strategy...	Generic > Circuit	p < .005
Manage my time efficiently...	Generic > Circuit	p < .05
Manage my time efficiently...	Choice > Circuit	p < .05
Write a program for a simple problem	Generic > Circuit	p < .005
Write a program for a simple problem	Choice > Circuit	p < .05
Write a program for a simple problem	Fantasy > Circuit	p < .05
Write a program for a moderately difficult problem	Generic > Circuit	p < .001
Write a program for a moderately difficult problem	Fantasy > Circuit	p < .05
Mentally trace through the execution of a long...	Generic > Circuit	p < .05
Mentally trace through the execution of a long...	Generic > Fantasy	p < .05
Write logically correct blocks of code	Generic > Choice	p < .05
Write logically correct blocks of code	Generic > Circuit	p < .05

Post hoc comparisons (LSD) revealed that participants in both the generic and choice conditions scored higher on “Complete a program if I had no help at all” than participants in the circuit condition, $p < .05$. Participants in the generic condition scored higher on “Write a program for an extremely difficult problem” than participants in the circuit condition, $p < .05$. Participants in the generic condition scored higher on “Come up with a suitable strategy for a given problem in a short time” than participants in the circuit condition, $p < .005$. Participants in both the generic and choice condition scored higher on “Manage my time efficiently if I had a pressing deadline on a problem” than participants in the circuit condition, $p < .05$. Participants in the generic, choice, and fantasy conditions scored higher on “Write a program for a simple problem” than participants in the circuit condition, $p < .05$. Participants in both the generic and fantasy condition scored higher on “Write a program for a moderately difficult problem” than participants in the circuit condition, $p < .05$. Participants in the generic condition scored higher on “Mentally trace through the execution of a long, complex, program given to me” than both the circuit and fantasy conditions, $p < .05$. Participants in the generic condition scored higher on “Write logically correct blocks of code” than participants in both the choice and circuit conditions, $p < .05$.

Table D.7: Self-Efficacy—Posthocs

Effect		Value	Hypothesis			Partial Eta	
			<i>F</i>	<i>df</i>	Error <i>df</i>	Sig.	Squared
Intercept	Pillai's Trace	.989	2477.999 ^a	42.000	1126.000	.000	.989
	Wilks' Lambda	.011	2477.999 ^a	42.000	1126.000	.000	.989
	Hotelling's Trace	92.430	2477.999 ^a	42.000	1126.000	.000	.989
	Roy's Largest Root	92.430	2477.999 ^a	42.000	1126.000	.000	.989
NumericCondition	Pillai's Trace	.233	2.265	126.000	3384.000	.000	.078
	Wilks' Lambda	.783	2.284	126.000	3374.672	.000	.078
	Hotelling's Trace	.258	2.303	126.000	3374.000	.000	.079
	Roy's Largest Root	.142	3.824 ^b	42.000	1128.000	.000	.125

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept + NumericCondition

Table D.8: GEQ—MANOVA Multivariate F-tests

Dependent Variable	Condition	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Answer.g7flow	Generic	3.293	.064	3.167	3.420
	Choice	3.267	.066	3.138	3.396
	Fantasy	3.260	.064	3.134	3.385
	Circuit	3.326	.067	3.195	3.457
Answer.h8flow	Generic	2.650	.071	2.510	2.790
	Choice	2.760	.073	2.617	2.903
	Fantasy	2.582	.071	2.443	2.722
	Circuit	2.892	.074	2.747	3.038
Answer.i9flow	Generic	2.710	.077	2.559	2.861
	Choice	2.781	.079	2.627	2.935
	Fantasy	2.740	.076	2.590	2.890
	Circuit	2.864	.080	2.707	3.020
Answer.j10flow	Generic	3.803	.061	3.684	3.923
	Choice	3.795	.062	3.674	3.917
	Fantasy	3.793	.060	3.674	3.911
	Circuit	3.871	.063	3.747	3.995
Answer.k11flow	Generic	2.377	.068	2.243	2.511
	Choice	2.403	.070	2.266	2.539
	Fantasy	2.303	.068	2.170	2.436
	Circuit	2.491	.071	2.352	2.630
Answer.l12flow	Generic	3.340	.069	3.205	3.475
	Choice	3.424	.070	3.286	3.561
	Fantasy	3.385	.068	3.251	3.519
	Circuit	3.459	.071	3.319	3.598
Answer.m13imm	Generic	2.093	.071	1.953	2.234
	Choice	2.066	.073	1.923	2.209
	Fantasy	2.039	.071	1.900	2.179
	Circuit	2.211	.074	2.066	2.357
Answer.n14imm	Generic	2.443	.061	2.323	2.564
	Choice	3.118	.063	2.995	3.241
	Fantasy	3.211	.061	3.091	3.330
	Circuit	3.086	.064	2.961	3.211
Answer.o15imm	Generic	2.833	.071	2.693	2.973
	Choice	2.927	.073	2.784	3.070
	Fantasy	3.069	.071	2.930	3.208
	Circuit	3.050	.074	2.905	3.195
Answer.p16imm	Generic	2.213	.073	2.071	2.356
	Choice	2.153	.074	2.007	2.298
	Fantasy	2.336	.072	2.194	2.477
	Circuit	2.534	.075	2.386	2.682
Answer.q17imm	Generic	2.780	.065	2.653	2.907
	Choice	2.983	.066	2.853	3.112
	Fantasy	2.993	.064	2.867	3.120
	Circuit	3.082	.067	2.951	3.214
Answer.r18imm	Generic	2.473	.066	2.344	2.603
	Choice	2.649	.068	2.517	2.782
	Fantasy	2.661	.066	2.532	2.790
	Circuit	2.749	.069	2.614	2.884
Answer.s19comp	Generic	3.187	.071	3.048	3.326
	Choice	3.247	.072	3.105	3.388

Table D.9: GEQ—Descriptives

	Fantasy	3.257	.070	3.118	3.395
	Circuit	3.115	.074	2.970	3.259
Answer.t20comp	Generic	2.170	.067	2.039	2.301
	Choice	2.375	.068	2.242	2.508
	Fantasy	2.362	.066	2.232	2.492
	Circuit	2.280	.069	2.144	2.415
Answer.u21comp	Generic	3.147	.065	3.019	3.274
	Choice	3.181	.066	3.051	3.310
	Fantasy	3.118	.064	2.992	3.245
	Circuit	2.867	.067	2.735	2.999
Answer.v22comp	Generic	3.227	.069	3.092	3.362
	Choice	3.226	.070	3.088	3.363
	Fantasy	3.211	.068	3.077	3.345
	Circuit	3.068	.071	2.928	3.208
Answer.w23comp	Generic	3.020	.064	2.894	3.146
	Choice	2.917	.066	2.788	3.045
	Fantasy	3.013	.064	2.888	3.138
	Circuit	2.649	.067	2.518	2.779
Answer.x24comp	Generic	3.317	.066	3.188	3.446
	Choice	3.309	.067	3.177	3.441
	Fantasy	3.375	.065	3.247	3.503
	Circuit	2.968	.068	2.834	3.102
Answer.y25chal	Generic	3.510	.069	3.375	3.645
	Choice	3.576	.070	3.439	3.714
	Fantasy	3.484	.068	3.349	3.618
	Circuit	3.613	.071	3.473	3.753
Answer.z26chal	Generic	2.980	.063	2.856	3.104
	Choice	3.003	.065	2.877	3.130
	Fantasy	3.072	.063	2.949	3.196
	Circuit	3.355	.066	3.226	3.484
Answer.za27chal	Generic	3.330	.065	3.202	3.458
	Choice	3.413	.067	3.282	3.544
	Fantasy	3.359	.065	3.231	3.486
	Circuit	3.487	.068	3.355	3.620
Answer.zb28chal	Generic	3.657	.062	3.535	3.778
	Choice	3.809	.063	3.685	3.933
	Fantasy	3.780	.062	3.659	3.900
	Circuit	3.961	.064	3.835	4.087
Answer.zc29chal	Generic	3.193	.065	3.066	3.321
	Choice	3.115	.066	2.984	3.245
	Fantasy	3.250	.065	3.123	3.377
	Circuit	3.423	.068	3.290	3.555
Answer.zd30chal	Generic	1.733	.063	1.610	1.856
	Choice	1.701	.064	1.576	1.827
	Fantasy	1.687	.062	1.565	1.810
	Circuit	1.943	.065	1.815	2.070
Answer.ze31tens	Generic	1.943	.065	1.816	2.070
	Choice	1.965	.066	1.836	2.095
	Fantasy	1.947	.064	1.821	2.073
	Circuit	2.262	.067	2.130	2.393
Answer.zf32tens	Generic	1.843	.059	1.727	1.960
	Choice	1.736	.061	1.617	1.855
	Fantasy	1.773	.059	1.657	1.889
	Circuit	1.996	.062	1.875	2.117
Answer.zg33tens	Generic	2.360	.074	2.215	2.505
	Choice	2.281	.076	2.133	2.430
	Fantasy	2.286	.074	2.142	2.431
	Circuit	2.520	.077	2.369	2.671

Answer.zh34tens	Generic	1.933	.068	1.800	2.067
	Choice	1.903	.069	1.766	2.039
	Fantasy	1.812	.068	1.680	1.945
	Circuit	2.090	.071	1.951	2.228
Answer.zi35tens	Generic	2.350	.076	2.200	2.500
	Choice	2.351	.078	2.198	2.503
	Fantasy	2.224	.076	2.075	2.372
	Circuit	2.545	.079	2.390	2.700
Answer.zj36tens	Generic	1.707	.060	1.589	1.824
	Choice	1.618	.061	1.498	1.738
	Fantasy	1.599	.059	1.482	1.715
	Circuit	1.914	.062	1.792	2.036
Answer.zk37pos	Generic	2.743	.061	2.625	2.862
	Choice	2.844	.062	2.723	2.965
	Fantasy	2.842	.060	2.724	2.960
	Circuit	2.742	.063	2.619	2.865
Answer.zl38pos	Generic	2.433	.069	2.299	2.568
	Choice	2.597	.070	2.460	2.735
	Fantasy	2.589	.068	2.455	2.722
	Circuit	2.448	.071	2.309	2.588
Answer.zm39pos	Generic	2.800	.062	2.679	2.921
	Choice	2.903	.063	2.779	3.026
	Fantasy	2.872	.061	2.752	2.992
	Circuit	2.889	.064	2.763	3.014
Answer.zn40pos	Generic	2.983	.061	2.864	3.103
	Choice	3.108	.062	2.985	3.230
	Fantasy	3.109	.061	2.990	3.228
	Circuit	3.079	.063	2.955	3.203
Answer.zo41pos	Generic	3.417	.065	3.290	3.543
	Choice	3.465	.066	3.336	3.594
	Fantasy	3.418	.064	3.292	3.544
	Circuit	3.444	.067	3.313	3.576
Answer.zp42pos	Generic	3.357	.067	3.225	3.489
	Choice	3.399	.069	3.264	3.534
	Fantasy	3.395	.067	3.264	3.526
	Circuit	3.387	.070	3.250	3.524
Answer.zq43neg	Generic	2.537	.065	2.409	2.664
	Choice	2.580	.066	2.449	2.710
	Fantasy	2.618	.065	2.491	2.745
	Circuit	2.437	.068	2.305	2.570
Answer.zr44neg	Generic	2.457	.071	2.317	2.596
	Choice	2.396	.073	2.253	2.538
	Fantasy	2.444	.071	2.305	2.583
	Circuit	2.538	.074	2.393	2.682
Answer.zs45neg	Generic	2.283	.071	2.144	2.422
	Choice	2.240	.072	2.098	2.381
	Fantasy	2.247	.070	2.109	2.385
	Circuit	2.308	.073	2.164	2.452
Answer.zt46neg	Generic	1.937	.058	1.822	2.051
	Choice	1.833	.059	1.717	1.950
	Fantasy	1.974	.058	1.860	2.087
	Circuit	1.896	.060	1.777	2.015
Answer.zu47neg	Generic	2.930	.081	2.771	3.089
	Choice	3.003	.083	2.841	3.165
	Fantasy	3.105	.080	2.948	3.263
	Circuit	2.910	.084	2.746	3.075
Answer.zv48neg	Generic	1.610	.055	1.503	1.717
	Choice	1.500	.056	1.391	1.609

Fantasy	1.513	.054	1.407	1.619
Circuit	1.699	.057	1.588	1.810

Dependent Variable	Conditions	p-value
"I forgot everything around me" (flow)	Circuit > Generic, Fantasy	p < .05
"It was aesthetically pleasing" (immersion)	Circuit, Fantasy, Choice > Generic	p < .001
"I felt that I could explore things" (immersion)	Circuit > Generic, Choice	p < .005
"I found it impressive" (immersion)	Circuit, Fantasy, Choice > Generic	p < .05
"It felt like a rich experience" (immersion)	Circuit, Fantasy > Generic	p < .05
"I was good at it" (competence)	Generic, Choice, Fantasy > Circuit	p < .01
"I was fast at reaching the game's targets" (competence)	Generic, Choice, Fantasy > Circuit	p < .005
"I felt competent" (competence)	Generic, Choice, Fantasy > Circuit	p < .001
"I thought it was hard" (challenge)	Circuit > Generic, Choice, Fantasy	p < .005
"I felt challenged" (challenge)	Circuit > Generic, Fantasy	p < .05
"I had to put a lot of effort into it" (challenge)	Circuit > Generic, Choice	p < .05
"I felt time pressure" (challenge)	Circuit > Generic, Choice, Fantasy	p < .05
"I felt tense" (tension)	Circuit > Generic, Choice, Fantasy	p < .005
"I felt restless" (tension)	Circuit > Choice, Fantasy	p < .01
"I felt irritable" (tension)	Circuit > Fantasy	p < .005
"I felt frustrated" (tension)	Circuit > Fantasy	p < .005
"I felt pressured" (tension)	Circuit > Generic, Choice, Fantasy	p < .05
"I felt tense" (tension)	Circuit > Generic, Choice, Fantasy	p < .005
"I felt tense" (tension)	Circuit > Generic, Choice, Fantasy	p < .005
"I felt tense" (tension)	Circuit > Generic, Choice, Fantasy	p < .005
"It gave me a bad mood" (negative affect)	Circuit > Fantasy, Choice	p < .05

Post hoc comparisons (LSD) revealed that participants in the circuit condition scored higher on "I forgot everything around me" (flow) than participants in the generic and fantasy conditions, $p < .05$. Participants in the circuit, fantasy, and choice conditions scored higher on "It was aesthetically pleasing" (immersion) than participants in the generic condition, $p < .001$. Participants in the circuit condition scored higher on "I felt that I could explore things" (immersion) than participants in the generic and choice conditions, $p < .005$. Participants in the circuit, fantasy, and choice conditions scored higher on "I found it impressive" (immersion) than participants in the generic condition, $p < .05$. Participants in the circuit and fantasy conditions scored higher on "It felt like a rich experience" (immersion) than participants in the generic condition, $p < .05$. Participants in the generic, choice, and fantasy conditions scored higher on "I was good at it" (competence) than participants in the circuit condition, $p < .01$. Participants in the generic, choice, and fantasy conditions scored higher on "I was fast at reaching the game's targets" (competence) than participants in the circuit condition, $p < .005$. Participants in the generic, choice, and fantasy conditions scored higher on "I felt competent" (competence) than participants in the circuit condition, $p < .001$. Participants in the circuit condition scored higher on "I thought it was hard" (challenge) than participants in the generic, choice, and fantasy conditions, $p < .005$. Participants in the circuit condition scored higher on "I felt challenged" (challenge) than

Table D.10: GEQ—Posthocs

participants in the generic and fantasy conditions, $p < .05$. Participants in the circuit condition scored higher on "I had to put a lot of effort into it" (challenge) than participants in the generic and choice conditions, $p < .05$. Participants in the circuit condition scored higher on "I felt time pressure" (challenge) than participants in the generic, choice, and fantasy conditions, $p < .05$. Participants in the circuit condition scored higher on "I felt tense" (tension) than participants in the generic, choice, and fantasy conditions, $p < .005$. Participants in the circuit condition scored higher on "I felt restless" (tension) than participants in the choice and fantasy conditions, $p < .01$. Participants in the circuit condition scored higher on "I felt irritable" (tension) than participants in the fantasy condition, $p < .005$. Participants in the circuit condition scored higher on "I felt frustrated" (tension) than participants in the fantasy condition, $p < .005$. Participants in the circuit condition scored higher on "I felt pressured" (tension) than participants in the generic, choice, and fantasy conditions, $p < .05$. Participants in the circuit condition scored higher on "It gave me a bad mood" (negative affect) than participants in the fantasy and choice conditions, $p < .05$.

Effect		Value	F	Hypothesis		Partial Eta	
				df	Error df	Sig.	Squared
Intercept	Pillai's Trace	.957	1226.218 ^a	21.000	1147.000	.000	.957
	Wilks' Lambda	.043	1226.218 ^a	21.000	1147.000	.000	.957
	Hotelling's Trace	22.450	1226.218 ^a	21.000	1147.000	.000	.957
	Roy's Largest Root	22.450	1226.218 ^a	21.000	1147.000	.000	.957
	Root						
NumericCondition	Pillai's Trace	.097	1.821	63.000	3447.000	.000	.032
	Wilks' Lambda	.905	1.840	63.000	3424.616	.000	.033
	Hotelling's Trace	.102	1.860	63.000	3437.000	.000	.033
	Roy's Largest Root	.075	4.104 ^b	21.000	1149.000	.000	.070
	Root						

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept + NumericCondition

Table D.11: PENS—MANOVA Multivariate F-tests

Dependent Variable	Condition	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Answer.zw49penscomp	Generic	4.540	.098	4.348	4.732
	Choice	4.486	.100	4.290	4.682
	Fantasy	4.503	.097	4.312	4.694
	Circuit	4.025	.102	3.826	4.225
Answer.zx50penscomp	Generic	4.413	.095	4.227	4.600
	Choice	4.392	.097	4.202	4.582
	Fantasy	4.312	.094	4.128	4.497
	Circuit	3.982	.098	3.789	4.175
Answer.zy51penscomp	Generic	4.437	.099	4.243	4.631
	Choice	4.507	.101	4.309	4.705
	Fantasy	4.444	.098	4.252	4.637
	Circuit	4.065	.102	3.864	4.266
Answer.zz52pensauton	Generic	3.903	.106	3.696	4.111
	Choice	3.986	.108	3.774	4.198
	Fantasy	4.049	.105	3.843	4.256
	Circuit	4.068	.110	3.853	4.283
Answer.zza53pensauton	Generic	3.723	.104	3.519	3.928
	Choice	3.941	.106	3.732	4.150
	Fantasy	3.862	.104	3.659	4.065
	Circuit	4.014	.108	3.802	4.226
Answer.zzb54pensauton	Generic	2.633	.093	2.451	2.816
	Choice	2.778	.095	2.591	2.964
	Fantasy	2.852	.092	2.671	3.033
	Circuit	3.186	.096	2.997	3.376
Answer.zzc55pensrelatedness	Generic	1.970	.082	1.810	2.130
	Choice	2.017	.083	1.854	2.181
	Fantasy	2.053	.081	1.893	2.212
	Circuit	2.111	.085	1.945	2.277
Answer.zzd56pensrelatedness	Generic	1.893	.080	1.737	2.050
	Choice	1.882	.082	1.722	2.042
	Fantasy	1.928	.079	1.772	2.083
	Circuit	2.093	.083	1.931	2.256
Answer.zze57pensrelatedness_rev	Generic	5.283	.130	5.028	5.539
	Choice	5.406	.133	5.145	5.667
	Fantasy	5.273	.130	5.019	5.527
	Circuit	5.057	.135	4.792	5.323
Answer.zzf58penspresence	Generic	2.127	.088	1.955	2.299
	Choice	2.170	.090	1.995	2.346
	Fantasy	2.296	.087	2.125	2.467
	Circuit	2.387	.091	2.209	2.566
Answer.zzg59penspresence	Generic	1.793	.080	1.637	1.950
	Choice	1.740	.081	1.580	1.899
	Fantasy	1.951	.079	1.795	2.106
	Circuit	2.054	.083	1.892	2.216
Answer.zzh60penspresence	Generic	1.733	.077	1.583	1.884
	Choice	1.771	.078	1.617	1.924
	Fantasy	1.855	.076	1.706	2.005
	Circuit	1.968	.080	1.812	2.124

Table D.12: PENS—Descriptives

Answer.zzi61penspresence_rev	Generic	5.073	.131	4.816	5.331
	Choice	4.858	.134	4.595	5.121
	Fantasy	4.914	.131	4.658	5.171
	Circuit	4.645	.136	4.378	4.913
Answer.zzj62penspresence	Generic	2.053	.083	1.891	2.215
	Choice	2.059	.084	1.894	2.224
	Fantasy	2.007	.082	1.846	2.167
	Circuit	2.201	.086	2.033	2.369
Answer.zzk63penspresence	Generic	1.517	.070	1.379	1.654
	Choice	1.601	.072	1.460	1.741
	Fantasy	1.553	.070	1.416	1.689
	Circuit	1.601	.072	1.460	1.741
Answer.zzl64penspresence	Generic	1.767	.078	1.614	1.919
	Choice	1.771	.079	1.615	1.927
	Fantasy	1.865	.077	1.714	2.017
	Circuit	1.814	.081	1.655	1.972
Answer.zzm65penspresence	Generic	3.873	.106	3.665	4.082
	Choice	3.969	.109	3.756	4.182
	Fantasy	3.980	.106	3.773	4.188
	Circuit	4.115	.110	3.898	4.331
Answer.zzn66penspresence	Generic	1.607	.068	1.472	1.741
	Choice	1.663	.070	1.526	1.800
	Fantasy	1.674	.068	1.541	1.808
	Circuit	1.796	.071	1.656	1.935
Answer.zzo67penscontrols	Generic	4.773	.102	4.574	4.973
	Choice	4.580	.104	4.376	4.784
	Fantasy	4.664	.101	4.466	4.863
	Circuit	4.699	.106	4.492	4.906
Answer.zzp68penscontrols	Generic	3.950	.104	3.746	4.154
	Choice	3.743	.106	3.535	3.951
	Fantasy	3.937	.103	3.735	4.140
	Circuit	4.032	.108	3.821	4.243
Answer.zzq69penscontrols	Generic	4.723	.103	4.521	4.926
	Choice	4.437	.105	4.231	4.644
	Fantasy	4.681	.103	4.480	4.882
	Circuit	4.613	.107	4.403	4.823

Dependent Variable	Conditions	p-value
PENS_competence_1	Circuit < Generic, Choice, Fantasy	p < .001
PENS_competence_2	Circuit < Generic, Choice, Fantasy	p < .05
PENS_competence_3	Circuit < Generic, Choice, Fantasy	p < .01
PENS_autonomy_3	Circuit > Generic, Choice, Fantasy	p < .05
PENS_presence_2	Circuit > Generic, Choice	p < .05
PENS_presence_6	Circuit > Generic, Choice, Fantasy	p < .005

Across all 3 questions on competence, participants in the circuit condition scored lower than participants in the generic, choice, and fantasy conditions, $p < .05$. For the question on autonomy, participants in the circuit condition scored higher than participants in the generic, choice, and fantasy conditions, $p < .05$. For one of the two questions on presence, participants in the circuit condition scored higher than participants in the generic, and choice conditions, $p < .05$. For the other of the two questions on presence, participants in the circuit condition scored higher than participants in the generic, choice and fantasy conditions, $p < .005$.

Table D.13: PENS—Posthocs

Dependent Variable	Choice	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Levels Completed	Generic	7.959	.374	7.222	8.696
	Fantasy	8.135	.263	7.618	8.653
	Circuit	7.791	.391	7.022	8.560
Hints Requested	Generic	1.699	.358	.995	2.402
	Fantasy	2.608	.251	2.114	3.102
	Circuit	2.925	.373	2.191	3.660
Attempts	Generic	21.562	4.124	13.444	29.679
	Fantasy	33.034	2.896	27.333	38.735
	Circuit	41.836	4.305	33.363	50.309

Table D.14: Choice Condition—Descriptives

Bibliography

- [1] Immersive VR Education. URL <http://immersivevreducation.com/>. 214
- [2] LightBot (<http://lightbot.com/>), 2008. 45
- [3] CodeCombat, 2016. URL <https://codecombat.com/>. 45, 172
- [4] S. Abramovich and P. Wardrip. *Impact of Badges on Motivation to Learn*. Routledge, 2016. 184, 188
- [5] S. Abramovich, C. Schunn, and R. M. Higashi. Are badges useful in education?: It depends upon the type of badge and expertise of learner. *Educational Technology Research and Development*, 61(2):217–232, 2013. ISSN 10421629. doi: 10.1007/s11423-013-9289-2. 187
- [6] L. C. Abroms, N. Padmanabhan, L. Thaweethai, and T. Phillips. iPhone apps for smoking cessation: A content analysis. *American Journal of Preventive Medicine*, 40(3):279–285, 2011. ISSN 07493797. doi: 10.1016/j.amepre.2010.10.032. 214
- [7] T. Agoritsas, E. Iserman, N. Hobson, N. Cohen, A. Cohen, P. S. Roshanov, M. Perez, C. Cotoi, R. Parrish, E. Pullenayegum, and Others. Increasing the quantity and quality of searching for current best evidence to answer clinical questions: protocol and intervention design of the MacPLUS FS Factorial Randomized Controlled Trials. *Implementation Science*, 9(1):125, 2014. 187

- [8] J. Ahn, A. Pellicone, and B. S. Butler. Open badges for education: What are the implications at the intersection of open systems and badging? *Research in Learning Technology*, 22(1063519):1–13, 2014. ISSN 21567077. doi: 10.3402/rlt.v22.23563. 186
- [9] A. V. Aho. Computation and computational thinking. *The Computer Journal*, 55(7): 832–835, 2012. 40
- [10] G. S. Aikenhead. *Science education for everyday life: Evidence-based practice*. Teachers College Press, 2006. 188
- [11] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Steering user behavior with badges. *WWW '13 Proceedings of the 22nd international conference on World Wide Web*, pages 95–106, 2013. doi: 10.1145/2488388.2488398. URL <http://dl.acm.org/citation.cfm?id=2488388.2488398>{%}5Cn<http://dl.acm.org/citation.cfm?id=2488398>. 184, 186, 187
- [12] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*, pages 687–698. ACM, 2014. 184, 187, 207
- [13] J. L. Andreacci, L. M. LeMura, S. L. Cohen, E. a. Urbansky, S. a. Chelland, and S. P. Von Duvillard. The effects of frequency of encouragement on performance during maximal exercise testing. *Journal of sports sciences*, 20(4):345–352, 2002. ISSN 0264-0414. doi: 10.1080/026404102753576125. 155, 164
- [14] S. Angel. MiiSearch, 2010. URL <http://www.miisearch.com/myavatareditor.swf>. 99
- [15] J. Antin and E. F. Churchill. Badges in social media: A social psychological perspective. In *CHI 2011 Gamification Workshop Proceedings*, pages 1–4. ACM New York, NY, 2011. 187
- [16] J. Aronson, M. J. Lustina, C. Good, K. Keough, C. M. Steele, and J. Brown. When

- white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of experimental social psychology*, 35(1):29–46, 1999. [33](#)
- [17] J. Aronson, C. B. Fried, and C. Good. Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, 2002. ISSN 00221031. doi: 10.1006/jesp.2001.1491. [34](#)
- [18] I. Arroyo, B. P. Woolf, J. M. Royer, and M. Tai. Affective gendered learning companions. In *Frontiers in Artificial Intelligence and Applications*, volume 200, pages 41–48, 2009. ISBN 9781607500285. doi: 10.3233/978-1-60750-028-5-41. [31](#), [132](#), [136](#)
- [19] M. Asgari and D. Kaufman. Relationships Among Computer Games, Fantasy, and Learning. *Proceedings of the 2nd International Conference on Imagination and Education*, pages 1–8, 2004. ISSN 0717-6163. doi: 10.1007/s13398-014-0173-7.2. [167](#), [172](#)
- [20] A. Assor, H. Kaplan, and G. Roth. Choice is good, but relevance is excellent: autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in schoolwork. *The British journal of Educational Psychology*, 72(Pt 2):261–278, 2002. ISSN 0007-0998. doi: 10.1348/000709902158883. URL <http://doi.wiley.com/10.1348/000709902158883>. [168](#)
- [21] C. M. Bachen, P. Hernández-Ramos, C. Raphael, and A. Waldron. How do presence, flow, and character identification affect players' empathy and interest in learning from a serious computer game? *Computers in Human Behavior*, 64:77–87, 2016. ISSN 07475632. doi: 10.1016/j.chb.2016.06.043. URL <http://dx.doi.org/10.1016/j.chb.2016.06.043>. [189](#)
- [22] J. N. Bailenson, J. Blascovich, and R. E. Guadagno. Self-representations in immersive virtual environments. *Journal of Applied Social Psychology*, 38(11):2673–2690, 2008. ISSN 00219029. [31](#), [132](#), [136](#), [184](#), [189](#)

- [23] R. Bailey, K. Wise, and P. Bolls. How Avatar Customizability Affects Children's Arousal and Subjective Presence During Junk Food-Sponsored Online Video Games. *CyberPsychology & Behavior*, pages 1–9, 2009. URL <http://online.liebertpub.com/doi/abs/10.1089/cpb.2008.0292>. 199
- [24] A. Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977. 79, 197
- [25] J. Banks and N. D. Bowman. Close intimate playthings? Understanding player-avatar relationships as a function of attachment, agency, and intimacy. *Selected Papers of Internet Research*, pages 1–4, 2013. 131, 132, 137
- [26] J. Banks and N. D. Bowman. Emotion, anthropomorphism, realism, control: Validation of a merged metric for player-avatar interaction (PAX). *Computers in Human Behavior*, 2016. 137
- [27] S. Barab and T. Duffy. From practice fields to communities of practice. *Theoretical foundations of learning ...*, 2000. URL <http://books.google.com/books?hl=en{&}lr={&}id=3oOpAgAAQBAJ{&}oi=fnd{&}pg=PA29{&}dq=From+Practice+Fields+to+Communities+of+Practice{&}ots=9m4u4cFE84{&}sig=89M5sTx4UmXx9B6ZN6HAhUaA6fo>. 167
- [28] V. Barr and C. Stephenson. Bringing computational thinking to K-12: what is Involved and what is the role of the computer science education community? *Acm Inroads*, 2(1):48–54, 2011. 43
- [29] C. Bauckhage and K. Kersting. How players lose interest in playing a game: An empirical study based on distributions of total playing times. *CIG*, pages 139–146, sep 2012. doi: 10.1109/CIG.2012.6374148. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6374148>http://ieeexplore.ieee.org/xpls/abs/_all.jsp?arnumber=6374148. 131, 136

- [30] R. F. Baumeister and M. R. Leary. The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin*, 117(3):497, 1995. 78, 197
- [31] R. F. Baumeister and D. M. Tice. Self-esteem and responses to success and failure: Subsequent performance and intrinsic motivation. *Journal of Personality*, 53(September 1985):450–467, 1985. ISSN 1467-6494. doi: 10.1111/j.1467-6494.1985.tb00376.x. 165
- [32] A. Baylor and Y. Kim. Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. *Intelligent Tutoring Systems*, (1997):592–603, 2004. ISSN 07413106. doi: 10.1007/978-3-540-30139-4_56. URL http://link.springer.com/chapter/10.1007/978-3-540-30139-4_{_}56. 31, 44, 130, 132, 136, 138, 184, 189
- [33] A. L. Baylor and S. J. Ebbers. Evidence that Multiple Agents Facilitate Greater Learning. *Artificial Intelligence in Education: Shaping the Future of Learning Through Intelligent Technologies*, pages 377–379, 2003. 137
- [34] A. L. Baylor and Y. Kim. Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, 15(1):95–115, 2005. ISSN 15604292. doi: 10.1007/BF02504991. URL http://digitalcommons.usu.edu/itls_{_}facpub/66/. 137
- [35] D. A. Becker. The Effects Of Choice On Auditors’ Intrinsic Motivation and Performance. *Behavioral Research in Accounting*, 9, 1997. ISSN 1050-4753. URL <http://search.ebscohost.com/login.aspx?direct=true{&}db=bth{&}AN=9708106803{&}site=ehost-live{&}scope=site>. 168
- [36] S. L. Beilock, W. A. Jellison, R. J. Rydell, A. R. McConnell, and T. H. Carr. On the causal mechanisms of stereotype threat: Can skills that don’t rely heavily on working memory still be threatened? *Personality and Social Psychology Bulletin*, 32(8):1059–1071, 2006. 34

- [37] R. Bellini, Y. Kleiman, and D. Cohen-Or. Time-varying Weathering in Texture Space. *SIGGRAPH*, 2016. ISSN 15577368. doi: 10.1145/2897824.2925891. 39
- [38] Y. Ben-Ari. Developing networks play a similar melody, 2001. ISSN 01662236.
- [39] B. M. Ben-David, P. H. H. M. van Lieshout, and T. Leszcz. A resource of validated affective and neutral sentences to assess identification of emotion in spoken language after a brain injury. *Brain injury : [BI]*, 25(2):206–20, 2011. ISSN 1362-301X. doi: 10.3109/02699052.2010.536197. URL <http://www.ncbi.nlm.nih.gov/pubmed/21117915>. 156
- [40] T. Ben-Zeev, S. Fein, and M. Inzlicht. Arousal and stereotype threat. *Journal of Experimental Social Psychology*, 41(2):174–181, 2005. ISSN 00221031. doi: 10.1016/j.jesp.2003.11.007. 34
- [41] A. J. Berinsky, G. A. Huber, and G. S. Lenz. Evaluating online labor markets for experimental research: Amazon. com’s Mechanical Turk. *Political Analysis*, 20(3): 351–368, 2012. 80
- [42] M. Berland, R. S. Baker, and P. Blikstein. Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, 19(1-2):205–220, 2014. ISSN 22111662. doi: 10.1007/s10758-014-9223-7. 38
- [43] P. C. Bernhardt, J. M. Dabbs, J. a. Fielden, and C. D. Lutter. Testosterone changes during vicarious experiences of winning and losing among fans at sporting events. *Physiology and Behavior*, 65(1):59–62, 1998. ISSN 00319384. doi: 10.1016/S0031-9384(98)00147-4. 165
- [44] K. Bessière, A. Seay, and S. Kiesler. The ideal elf: Identity exploration in World of Warcraft. *CyberPsychology & Behavior*, 2007. URL <http://online.liebertpub.com/doi/abs/10.1089/cpb.2007.9994>. 189, 207
- [45] N. E. Betz and G. Hackett. Applications of Self-Efficacy Theory to Understanding

- Career Choice Behavior. *Journal of Social and Clinical Psychology*, 4(3):279–289, 1986. ISSN 0736-7236. doi: 10.1521/jscp.1986.4.3.279. 77, 172
- [46] M. V. Birk, R. L. Mandryk, M. K. Miller, and K. M. Gerling. How self-esteem shapes our interactions with play technologies. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, pages 35–45. ACM, 2015.
- [47] M. V. Birk, C. Atkins, J. T. Bowey, and R. L. Mandryk. Fostering Intrinsic Motivation through Avatar Identification in Digital Games. *CHI*, 2016. doi: 10.1145/2858036.2858062. 32, 184, 189, 190, 199, 208
- [48] M. V. Birk, R. L. Mandryk, and C. Atkins. The Motivational Push of Games. *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*, (April 2017):291–303, 2016. doi: 10.1145/2967934.2968091. URL <http://hci.usask.ca/uploads/396-avatarCHIPlay2016{ }camera{ }ready.pdf{ }0Ahttp://dl.acm.org/citation.cfm?doid=2967934.2968091>. 184, 189
- [49] S. K. Bista, S. Nepal, N. Colineau, and C. Paris. Using gamification in an online community. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on*, pages 611–618. IEEE, 2012.
- [50] S. K. Bista, S. Nepal, and C. Paris. Engagement and Cooperation in Social Networks: Do Benefits and Rewards Help? In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, pages 1405–1410. IEEE, 2012. 186
- [51] L. Blair. What Video Games Can Teach Us About Badges and Pathways. *Digital Badges in Education: Trends, Issues, and Cases*, page 62, 2016. 187
- [52] H. Blanton, J. Crocker, and D. T. Miller. The Effects of In-Group versus Out-Group Social Comparison on Self-Esteem in the Context of a Negative Stereotype. *Journal*

- of Experimental Social Psychology*, 36(5):519–530, sep 2000. ISSN 00221031. doi: 10.1006/jesp.2000.1425. URL <http://linkinghub.elsevier.com/retrieve/pii/S0022103100914252>. 122
- [53] J. Blascovich, S. J. Spencer, D. Quinn, and C. Steele. African Americans and high blood pressure: The role of stereotype threat. *Psychological science*, 12(3):225–229, 2001. 34
- [54] P. Blikstein. Using learning analytics to assess students’ behavior in open-ended programming tasks. In *Proceedings of the 1st international conference on learning analytics and knowledge*, pages 110–116. ACM, 2011. 38
- [55] P. Blikstein. Digital Fabrication and ‘Making’ in Education: The Democratization of Invention. *FabLabs: Of Machines, Makers and Inventors*, pages 1–21, 2013. ISSN 1074-9039. doi: 10.1080/10749039.2014.939762. URL <https://tltl.stanford.edu/sites/default/files/files/documents/publications/2013.Book-B.Digital.pdf>. 37, 38
- [56] P. C. Blumenfeld, T. M. Kempler, and J. S. Krajcik. Motivation and Cognitive Engagement in Learning Environments. In *The Cambridge Handbook of the Learning Sciences*, pages 475–488. 2005. ISBN 9780511816833. doi: <http://dx.doi.org/10.1017/CBO9780511816833.029>. URL <http://ebooks.cambridge.org/chapter.jsf?bid=CBO9780511816833&cid=CBO9780511816833A040&tabName=Chapter>. 20, 77, 123, 131, 136, 164, 168, 172
- [57] I. Bogost. Asynchronous multiplayer: Futures for casual multiplayer experience. *Other Players*, 2004. URL <http://www.bogost.com/downloads/I.Bogost-AsynchronousMultiplay.pdf>. 141
- [58] I. Bogost. Procedural literacy: Problem solving with programming, systems, and play. *Journal of Media Literacy*, pages 32–36, 2005. URL <http://scholar.google.com/scholar?hl=en&btnG=>

[Search{&}q=intitle:Procedural+Literacy+:+Problem+Solving+with+Programming+,+Systems+,+{&}+Play{#}0. 40](#)

- [59] I. Bogost. Gamification is bullshit. *The gameful world: Approaches, issues, applications*, pages 65–80, 2011. 186
- [60] I. Bogost. Exploitationware. In *Rhetoric/composition/play through video games*, pages 139–147. Springer, 2013. 186
- [61] A. Booth, G. Shelley, A. Mazur, G. Tharp, and R. Kittok. Testosterone, and winning and losing in human competition. *Hormones and behavior*, 23(4):556–571, 1989. ISSN 0018-506X. doi: 10.1016/0018-506X(89)90042-1. URL <http://www.sciencedirect.com/science/article/pii/S0018506X89900421>. 165
- [62] I. Boticki, J. Baksa, P. Seow, and C.-K. Looi. Usage of a mobile social learning platform with virtual badges in a primary school. *Computers & Education*, 86:120–136, 2015. 187
- [63] N. D. Bowman, R. Rogers, and B. I. Sherrick. In Control or In Their Shoes? Video games, characters, and enjoyable or meaningful experiences. *Broadcast Education Association Research Symposium "Media and Social Life: The Self, Relationships, and Society."*, 2013. 108, 131, 132, 137
- [64] N. D. Bowman, M. B. Oliver, R. Rogers, B. Sherrick, J. Woolley, and M.-Y. Chung. In control or in their shoes? How character attachment differentially influences video game enjoyment and appreciation. *Journal of Gaming & Virtual Worlds*, 8(1):83–99, 2016. ISSN 1757191X. doi: 10.1386/jgvw.8.1.83_1. URL <http://openurl.ingenta.com/content/xref?genre=article{&}issn=1757-191X{&}volume=8{&}issue=1{&}spage=83>. 189
- [65] P. Brauner, T. Leonhardt, M. Ziefle, and U. Schroeder. The effect of tangible artifacts, gender and subjective technical competence on teaching programming to seventh

- graders. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5941 LNCS, pages 61–71, 2010. ISBN 3642113753. doi: 10.1007/978-3-642-11376-5_7. 172
- [66] J. W. Brehm. *A theory of psychological reactance*. 1966. ISBN 0121298507. doi: 10.1002/hrdq.20027. URL <http://books.google.com/books?hl=en&lr=&id=JZ0rkeNvVkcC&oi=fnd&pg=PA377&dq=Theory+of+Psychological+Reactance&ots=nNrmO9ZUFe&sig=zEfZUDHNWnn21G5Qpp{ }0ifb-Cgk>. 132
- [67] K. Brennan and M. Resnick. New frameworks for studying and assessing the development of computational thinking. *Annual American Educational Research Association meeting, Vancouver, BC, Canada*, pages 1–25, 2012. 19, 39, 41, 190
- [68] S. P. Brown, W. L. Cron, and T. W. Leigh. Do feelings of success mediate sales performance-work attitude relationships? *Journal of the Academy of Marketing Science*, 21(2):91–100, 1993. ISSN 0092-0703. doi: 10.1007/BF02894420. URL <http://link.springer.com/10.1007/BF02894420>. 165
- [69] A. Bruckman, M. Biggers, and B. Ericson. "Georgia Computes!": Improving the Entire Computing Education Pipeline. *SIGCSE '09 Proceedings of the 40th ACM technical symposium on Computer science education*, 2009. URL <https://home.cc.gatech.edu/guzdial/uploads/170/GaComputes-SIGCSE2009-v5.docx>. 3
- [70] E. Brummelman, S. Thomaes, B. Orobio de Castro, G. Overbeek, and B. J. Bushman. "That's Not Just Beautiful—That's Incredibly Beautiful!": The Adverse Impact of Inflated Praise on Children With Low Self-Esteem. *Psychological Science*, 25(3): 728–735, 2014. ISSN 0956-7976. doi: 10.1177/0956797613514251. URL <http://pss.sagepub.com/lookup/doi/10.1177/0956797613514251>. 164
- [71] E. Brummelman, S. Thomaes, G. Overbeek, B. Orobio de Castro, M. a. van den Hout, and B. J. Bushman. On feeding those hungry for praise: Person praise backfires in

- children with low self-esteem. *Journal of Experimental Psychology: General*, 143(1): 9–14, 2014. ISSN 1939-2222. doi: 10.1037/a0031917. URL <http://www.ncbi.nlm.nih.gov/pubmed/23421441>. 156
- [72] J. C. Brunstein and P. M. Gollwitzer. Effects of failure on subsequent performance: the importance of self-defining goals. *Journal of personality and social psychology*, 70(2):395–407, 1996. ISSN 0022-3514. doi: 10.1037/0022-3514.70.2.395. 165
- [73] E. Bubl, E. Kern, D. Ebert, M. Bach, and L. Tebartz Van Elst. Seeing gray when feeling blue? Depression can be measured in the eye of the diseased. *Biological Psychiatry*, 68(2):205–208, 2010. ISSN 00063223. doi: 10.1016/j.biopsych.2010.02.009. URL <http://dx.doi.org/10.1016/j.biopsych.2010.02.009>. 147
- [74] L. Buechley. A construction kit for electronic textiles. In *Wearable Computers, 2006 10th IEEE International Symposium on*, pages 83–90. IEEE, 2006. 38
- [75] L. Buechley, M. Eisenberg, and J. Catchen. The LilyPad Arduino: Using computational textiles to investigate engagement, aesthetics, and diversity in computer science education. *CHI '08 Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 423–432, 2008. 38, 39, 45
- [76] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011. 80
- [77] K. R. Butcher. Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology*, 98(1):182–197, 2006. ISSN 0022-0663. doi: 10.1037/0022-0663.98.1.182. 173
- [78] A. P. Buunk, J. M. Peiró, and C. Griffioen. A positive role model may stimulate career-oriented behavior. *Journal of Applied Social Psychology*, 37:1489–1500, 2007.

- ISSN 00219029. doi: 10.1111/j.1559-1816.2007.00223.x. 34, 35, 129, 184, 188, 192, 209
- [79] D. Byrne and D. Nelson. Attraction as a linear function of proportion of positive reinforcements. *Journal of Personality and Social Psychology*, 1(6):659–663, 1965. ISSN 1939-1315. doi: 10.1037/h0022073. 31, 132, 136, 189
- [80] J. T. Cacioppo, W. L. Gardner, and G. G. Berntson. The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology*, 76(5):839–855, 1999. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.3238>. 147
- [81] J. A. Cafazzo, M. Casselman, N. Hamming, D. K. Katzman, and M. R. Palmert. Design of an mHealth app for the self-management of adolescent type 1 diabetes: a pilot study. *Journal of medical Internet research*, 14(3), 2012.
- [82] C. Calhoun. *Robert K. Merton: sociology of science and sociology as science*. Columbia University Press, 2010. 188
- [83] W. K. Campbell and C. Sedikides. Self-threat magnifies the self-serving bias: A meta-analytic integration., 1999. ISSN 1089-2680. URL <http://psycnet.apa.org/journals/gpr/3/1/23/>. 31, 131, 132, 136, 189
- [84] J. Cassell. Embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78, 2000. ISSN 00010782. doi: 10.1145/332051.332075. URL <http://dl.acm.org/citation.cfm?id=332075>. 27
- [85] J. Cassell and K. R. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538, 1999. 27
- [86] M. Chai, C. Zheng, and K. Zhou. A Reduced Model for Interactive Hairs. *ACM Trans. Graph.*, 33(4):124:1–124:11, 2014. ISSN 0730-0301. doi: 10.1145/2601097.2601211. URL <http://doi.acm.org/10.1145/2601097.2601211>. 39

- [87] M. Chai, T. Shao, H. Wu, Y. Weng, and K. Zhou. AutoHair: fully automatic hair modeling from a single image. *ACM Transactions on Graphics (TOG)*, 35(4):116, 2016. 39
- [88] J. Chandler and D. Shapiro. Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, 12, 2016. 80
- [89] J. Chandler, P. Mueller, and G. Paolacci. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46(1):112–130, 2014. ISSN 1554-3528. doi: 10.3758/s13428-013-0365-7. URL <http://link.springer.com/10.3758/s13428-013-0365-7>. 80
- [90] Y. H. Chang, Y. Y. Chen, N. S. Chen, Y. T. Lu, and R. J. Fang. Yet another adaptive learning management system based on Felder and Silverman’s Learning Styles and Mashup. *Eurasia Journal of Mathematics, Science and Technology Education*, 12(5): 1273–1285, 2016. ISSN 13058223. doi: 10.12973/eurasia.2016.1512a. 44
- [91] M. a. Changizi, Q. Zhang, and S. Shimojo. Bare skin, blood and the evolution of primate colour vision. *Biology letters*, 2(2):217–221, 2006. ISSN 1744-9561. doi: 10.1098/rsbl.2006.0440. 147
- [92] C. Chase. *Motivating Persistence in the Face of Failure: The Impact of an Ego-protective Buffer on Learning Choices and Outcomes in a Computer-based Educational Game*. Stanford University, 2011. URL [https://books.google.com/books?hl=en{&}lr={&}id=uHsdh6mKHuYC{&}pgis=1](https://books.google.com/books?hl=en&lr={&}id=uHsdh6mKHuYC{&}pgis=1). 165
- [93] S. Cheryan, B. J. Drury, and M. Vichayapai. Enduring Influence of Stereotypical Computer Science Role Models on Women’s Academic Aspirations. *Psychology of Women Quarterly*, 37(1):72–79, sep 2012. ISSN 0361-6843. doi: 10.1177/0361684312459328. URL <http://pwq.sagepub.com/lookup/doi/10.1177/0361684312459328>. 34

- [94] V. Chirkov, R. M. Ryan, Y. Kim, and U. Kaplan. Differentiating autonomy from individualism and independence: a self-determination theory perspective on internalization of cultural orientations and well-being. *Journal of personality and social psychology*, 84(1):97, 2003. 78, 197
- [95] K. Christoph, H. Dorothée, and V. Peter. The Video Game Experience as "True" Identification: A Theory of Enjoyable Alterations of Players' Self-Perception. *Communication theory*, 19(4):351–373, 2009. 188
- [96] P. W. Clark, C. A. Martin, and A. J. Bush. The effect of role model influence on adolescents' materialism and marketplace knowledge. *Journal of Marketing Theory and Practice*, 9(4):27–36, 2001. 208
- [97] R. C. Clark and R. E. Mayer. E-learning and the Science of Instruction : Proven Guidelines for Consumers and Designers of Multimedia Learning. 2011. 167
- [98] K. Claypool and M. Claypool. Teaching software engineering through game design. In *ACM SIGCSE Bulletin*, volume 37, pages 123–127. ACM, 2005. 45
- [99] Code.org. Code.org, 2014. URL <http://code.org>. 45
- [100] G. L. Cohen, J. Garcia, N. Apfel, and A. Master. Reducing the racial achievement gap: a social-psychological intervention. *Science (New York, N.Y.)*, 313(5791):1307–1310, 2006. ISSN 0036-8075. doi: 10.1126/science.1128317. 34
- [101] J. Cohen. Defining identification: A theoretical look at the identification of audiences with media characters. *Mass communication & society*, 4(3):245–264, 2001. 188, 189
- [102] J. Cohen. Deconstructing Ally: Explaining viewers' interpretations of popular television. *Media Psychology*, 4(3):253–277, 2002. 189
- [103] College Board. AP Computer Science Principles. URL <https://>

[//advancesinap.collegeboard.org/stem/computer-science-principles](http://advancesinap.collegeboard.org/stem/computer-science-principles). 40

- [104] Conrad's Home. Mii Editor, 2017. URL <http://www.conradshome.com/mii/>. 199
- [105] S. Cooper, W. Dann, and R. Pausch. Alice : a 3-D Tool for Introductory Programming Concepts. *Journal of Computing Sciences in Colleges*, 15(5):107–116, 2000. 38, 45
- [106] D. I. Cordova and M. R. Lepper. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88(4):715–730, 1996. ISSN 1939-2176(Electronic);0022-0663(Print). doi: 10.1037/0022-0663.88.4.715. 167, 168, 172, 188
- [107] N. R. Council and Others. *Report of a workshop on the pedagogical aspects of computational thinking*. National Academies Press, 2011. 43
- [108] H. Cramer, M. Rost, and L. E. Holmquist. Performing a check-in: emerging practices, norms and 'conflicts' in location-sharing using foursquare. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, pages 57–66. ACM, 2011.
- [109] J.-C. Croizet and T. Claire. Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, 24(6):588–594, 1998. 33
- [110] J.-C. Croizet, G. Després, M.-E. Gauzins, P. Huguet, J.-P. Leyens, and A. Méot. Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin*, 30(6):721–731, 2004. 34
- [111] K. Crowley and M. Jacobs. Building Islands of Expertise in Everyday Family Activity. *Learning Conversations in Museums*, pages 1–23, 2002. 37, 167
- [112] R. J. Czaja and R. G. Cummings. Designing Competitions : How To Maintain

- Motivation For Losers. *American Journal of Business Education*, 2(9):91–99, 2009. 165
- [113] H. Daanen and L. Grant. Space Mission: Ice Moon. In *ACM SIGGRAPH 2007 educators program on - SIGGRAPH '07*, page 19, New York, New York, USA, aug 2007. ACM Press. ISBN 9781450318303. doi: 10.1145/1282040.1282060. URL <http://dl.acm.org/citation.cfm?id=1282040.1282060>. 167, 172
- [114] Y. A. W. De Kort, W. A. IJsselsteijn, and K. Poels. Digital games as social presence technology: Development of the Social Presence in Gaming Questionnaire (SPGQ). *Proceedings of PRESENCE*, 195203, 2007. 78
- [115] E. Deci and R. M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum Press, 1985. 168
- [116] E. L. Deci and R. M. Ryan. Motivation, personality, and development within embedded social contexts: An overview of self-determination theory. *The Oxford handbook of human motivation*, pages 85–107, 2012. 78, 197
- [117] E. L. Deci, R. Koestner, and R. M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation., 1999. 186
- [118] A. R. Denham and K. W. Guyotte. Cultivating critical game makers in digital game-based learning: learning from the arts. *Learning, Media and Technology*, 9884(July):1–11, 2017. ISSN 1743-9884. doi: 10.1080/17439884.2017.1342655. URL <https://www.tandfonline.com/doi/full/10.1080/17439884.2017.1342655>. 190, 208, 209
- [119] A. Denisova and P. Cairns. The Placebo Effect in Digital Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*, pages 23–33, 2015. ISBN 9781450334662. doi: 10.1145/2793107.2793109. URL <http://dl.acm.org/citation.cfm?doid=2793107.2793109>. 165

- [120] P. J. Denning and P. A. Freeman. The profession of IT computing's paradigm. *Communications of the ACM*, 52(12):28–30, 2009. 43
- [121] P. Denny. The effect of virtual achievements on student engagement. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, page 763, 2013. ISSN 9781450318990. doi: 10.1145/2470654.2470763. URL <http://www.scopus.com/inward/record.url?eid=2-s2.0-84877936824&partnerID=tZ0tx3y1> <http://dl.acm.org/citation.cfm?id=2470654.2470763> <http://dl.acm.org/citation.cfm?doid=2470654.2470763>. 184, 186, 187, 207
- [122] S. Deterding. Situated motivational affordances of game elements: A conceptual model. In *Gamification: Using game design elements in non-gaming contexts, a workshop at CHI*, 2011. 186
- [123] S. Deterding. Coding conduct: Games, play, and human conduct between technical artifacts and social framing. 2012. URL <https://www.slideshare.net/dings/coding-conduct-games-play-and-human-conduct-between-technical-code-and-social-framing>. 186
- [124] S. Deterding. The Lens of Intrinsic Skill Atoms: A Method for Gameful Design. *Human-Computer Interaction*, 30(3-4):294–335, 2015. ISSN 0737-0024. doi: 10.1080/07370024.2014.993471. 78, 197
- [125] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From game design elements to gamefulness. *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11*, pages 9–11, 2011. ISSN 1450308163. doi: 10.1145/2181037.2181040. URL <http://doi.acm.org/10.1145/2181037.2181040> <http://dl.acm.org/citation.cfm?doid=2181037.2181040>. 186
- [126] V. Devedžić and J. Jovanović. Developing Open Badges: A comprehensive approach.

- Educational Technology Research and Development*, 63(4):603–620, 2015. ISSN 15566501. doi: 10.1007/s11423-015-9388-3.
- [127] J. Dewey. Interest and effort in education. *Search*, pages 1859–1952, 1913. 167
- [128] J. M. Dias Neto, F. B. Silva, A. L. B. D. Oliveira, N. L. Couto, E. H. M. Dantas, and M. A. D. L. Nascimento. Effects of verbal encouragement on performance of the multistage 20 m shuttle run. *Acta Scientiarum. Health Sciences*, 37(1):25, 2015. ISSN 1807-8648. doi: 10.4025/actascihealthsci.v37i1.23262. URL <http://periodicos.uem.br/ojs/index.php/ActaSciHealthSci/article/view/23262>. 155, 164
- [129] D. Dicheva, C. Dichev, G. Agre, and G. Angelova. Gamification in Education: A Systematic Mapping Study. *Educational Technology & Society*, 18(3):75–88, 2015. ISSN 1436-4522. doi: 10.1109/EDUCON.2014.6826129. 186, 187
- [130] B. J. DiSalvo, K. Crowley, and R. Norwood. Learning in Context: Digital Games and Young Black Men. *Games and Culture*, 3(2):131–141, feb 2008. ISSN 1555-4120. doi: 10.1177/1555412008314130. URL <http://gac.sagepub.com/cgi/doi/10.1177/1555412008314130>. 3
- [131] A. DiSessa. Changing Minds: Computers, Learning, and Literacy. *Journal of Teacher Education*, 60(1):38–51, 2000. ISSN 00224871. doi: Article. URL <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20{&}path=ASIN/0262041804>. 40
- [132] A. A. DiSessa. *Changing Minds: Computers, Learning, and Literacy*. MIT Press, 2001. ISBN 0262541327. URL <http://books.google.com/books?hl=en{&}lr={&}id=DfNaW4zvJVgC{&}pgis=1>. 3, 19
- [133] A. Domínguez, J. Saenz-De-Navarrete, L. De-Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz. Gamifying learning experiences: Practical implications and outcomes. *Computers & Education*, 63:380–392, 2013. 186, 187

- [134] I. X. Domínguez and D. L. Roberts. Asymmetric Virtual Environments : Exploring the Effects of Avatar Colors on Performance. *Experimental Artificial Intelligence in Games: Papers from the AIIDE Workshop*, 2015. 36, 148
- [135] P. Dourish. Seeking a Foundation for Context-Aware Computing. *Human-Computer Interaction*, 16:229–241, 2001. ISSN 07370024. doi: 10.1207/S15327051HCI16234_07. URL <http://portal.acm.org/citation.cfm?id=1463108.1463115>.
- [136] D. Dreiskaemper, B. Strauss, N. Hagemann, and D. Büsch. Influence of red jersey color on physical parameters in combat sports. *Journal of sport & exercise psychology*, 35(FEBRUARY):44–9, 2013. ISSN 1543-2904. URL <http://www.ncbi.nlm.nih.gov/pubmed/23404878>. 147, 148
- [137] J. E. Driskell and D. J. Dwyer. Microcomputer Videogame Based Training. *Educational Technology*, 24:11–17, 1984. 167
- [138] R. Driver, H. Asoko, J. Leach, P. Scott, and E. Mortimer. Constructing scientific knowledge in the classroom. *Educational researcher*, 23(7):5–12, 1994. 184, 188
- [139] B. J. Drury, J. O. Siy, and S. Cheryan. When Do Female Role Models Benefit Women? The Importance of Differentiating Recruitment From Retention in STEM. *Psychological Inquiry*, 22(4):265–269, oct 2011. ISSN 1047-840X. doi: 10.1080/1047840X.2011.620935. URL <http://www.tandfonline.com/doi/abs/10.1080/1047840X.2011.620935>. 34
- [140] N. Ducheneaut, M. Wen, N. Yee, and G. Wadley. Body and mind: a study of avatar personalization in three virtual worlds. *CHI 2009*, 2009. URL <http://dl.acm.org/citation.cfm?id=1518877>. 189, 207, 214
- [141] T. S. Duval and P. J. Silvia. Self-awareness, probability of improvement, and the self-serving bias. *Journal of personality and social psychology*, 82(1):49–61, 2002. ISSN 0022-3514. doi: 10.1037/0022-3514.82.1.49. 31, 131, 132, 136, 189

- [142] C. Dweck. *Mindset: The New Psychology of Success*. Random House Publishing Group, 2006. ISBN 1588365239. URL <http://books.google.com/books?hl=en&lr=&id=fdjqz0TPL2wC&pgis=1>. 155
- [143] A. J. Elliot and H. Aarts. Perception of the color red enhances the force and velocity of motor output. *Emotion (Washington, D.C.)*, 11(2):445–449, 2011. ISSN 1528-3542. doi: 10.1037/a0022599. 154
- [144] A. J. Elliot and M. V. Covington. Approach and avoidance motivation. *Educational Psychology Review*, 13(2):73–92, 2001. ISSN 1040726X. doi: 10.1023/A:1009009018235. 147
- [145] A. J. Elliot and M. a. Maier. Color-in-Context Theory. In *Advances in Experimental Social Psychology*, volume 45, pages 61–125. 2012. ISBN 9780123942869. doi: 10.1016/B978-0-12-394286-9.00002-0. URL <http://dx.doi.org/10.1016/B978-0-12-394286-9.00002-0>. 147, 153
- [146] A. J. Elliot, M. a. Maier, A. C. Moller, R. Friedman, and J. Meinhardt. Color and psychological functioning: the effect of red on performance attainment. *Journal of experimental psychology. General*, 136(1):154–168, 2007. ISSN 0096-3445. doi: 10.1037/0096-3445.136.1.154. 25, 93, 147, 148, 154, 193
- [147] R. Elliott, B. J. Sahakian, a. P. McKay, J. J. Herrod, T. W. Robbins, and E. S. Paykel. Neuropsychological impairments in unipolar depression: the influence of perceived failure on subsequent performance. *Psychological medicine*, 26:975–989, 1996. ISSN 0033-2917. doi: 10.1017/S0033291700035303. 165
- [148] Entertainment Software Association. Sales, Demographic, and Usage Data. 4(1): 2–4, 2015. ISSN 0894-4393. URL http://www.theesa.com/facts/pdfs/ESA{}_EF{}_2008.pdf. 19, 172
- [149] Y. Eshel and R. Kohavi. Perceived Classroom Control, Self-Regulated Learning Strategies, and Academic Achievement. *Educational Psychology*, 23(3):249,

2003. ISSN 1469-5820. doi: 10.1080/0144341032000060093. URL <http://search.ebscohost.com/login.aspx?direct=true{%&db=eue{%&AN=9428653{%&lang=it{%&site=ehost-live{%}%}5CnEshel,Kohavi2003-PerceivedClassroomControl.pdf>. 168
- [150] M. Evans and A. R. Boucher. Optimizing the power of choice: Supporting student autonomy to foster motivation and engagement in learning. *Mind, Brain, and Education*, 9(2):87–91, 2015. ISSN 1751228X. doi: 10.1111/mbe.12073. 171, 207
- [151] M. D. Fairchild. The Munsell book of color. In *Color Appearance Models*. 2005. 149
- [152] N. J. G. Falkner and K. E. Falkner. "Whither, Badges?" or "Wither, Badges!": A Metastudy of Badges in Computer Science Education to Clarify Effects , Significance and Influence. *Computer and Education*, 50(1): 127–135, 2013. ISSN 15508390. doi: 10.1002/meet.14505001101. URL <http://www.scopus.com/inward/record.url?eid=2-s2.0-84905457954{%&partnerID=tZ0tx3y1{%&}}5Cnhttp://dl.acm.org/citation.cfm?id=2583008.2583017>. 186
- [153] R. Farzan, J. M. DiMicco, D. R. Millen, C. Dugan, W. Geyer, and E. A. Brownholtz. Results from deploying a participation incentive mechanism within the enterprise. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 563–572. ACM, 2008. 186
- [154] G. Fauconnier and M. Turner. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books, 2008. 31
- [155] N. T. Feather. Effects of prior success and failure on expectations of success and subsequent performance. *Journal of personality and social psychology*, 3(3):287–98, 1966. ISSN 0022-3514. doi: <adata-auto="link"href="javascript:__doLinkPostBack('','ss~~DItitle="Searchfor10.1037/h0022965" id="link10.1037h0022965">10.1037/h0022965. URL <http://www.ncbi.nlm.nih.gov/pubmed/5906331>. 165

- [156] N. T. Feather. Attribution of responsibility and valence of success and failure in relation to initial confidence and task performance. *Journal of Personality and Social Psychology*, 13(2):129–144, 1969. ISSN 0022-3514. doi: 10.1037/h0028071. 165
- [157] C. V. Feilitzen and O. Linné. Identifying with television characters. *Journal of Communication*, 25(4):51–55, 1975. 207
- [158] R. Felder and L. Silverman. Learning and teaching styles in engineering education. *Engineering education*, 78(June):674–681, 1988. ISSN 01905848. doi: 10.1109/FIE.2008.4720326. URL <http://www.academia.edu/download/31039406/LS-1988.pdf>. 44
- [159] E. Fennema and J. A. Sherman. Fennema-Sherman mathematics attitudes scales: Instruments designed to measure attitudes toward the learning of mathematics by females and males. *Journal for research in Mathematics Education*, 7(5):324–326, 1976.
- [160] M. Filsecker and D. T. Hickey. A multilevel analysis of the effects of external rewards on elementary students’ motivation, engagement and learning in an educational game. *Computers and Education*, 75(August):136–148, 2014. ISSN 03601315. doi: 10.1016/j.compedu.2014.02.008. URL <http://dx.doi.org/10.1016/j.compedu.2014.02.008>. 187
- [161] J. Fišer, O. Jamriška, M. Lukáč, E. Shechtman, P. Asente, J. Lu, and D. Skora. StyLit: Illumination-Guided Example-Based Stylization of 3D Renderings. *SIGGRAPH*, 2016. ISSN 15577368. doi: 10.1145/2897824.2925948. 39
- [162] Z. Fitz-Walter, D. Tjondronegoro, and P. Wyeth. Orientation passport: using gamification to engage university students. In *Proceedings of the 23rd Australian computer-human interaction conference*, pages 122–125. ACM, 2011. 187
- [163] P. C. Flore and J. M. Wicherts. Does stereotype threat influence performance of girls

- in stereotyped domains? A meta-analysis. *Journal of school psychology*, 53(1):25–44, 2015. 35
- [164] T. Flowerday and G. Schraw. Teacher beliefs about instructional choice: A phenomenological study. *Journal of Educational Psychology*, 92(4):634–645, 2000. ISSN 0022-0663. doi: 10.1037/0022-0663.92.4.634. 168, 171, 207
- [165] A. Foster. Games and Motivation to Learn Science: Personal Identity, Applicability, Relevance and Meaningfulness. *Journal of Interactive Learning Research*, 19(4):597–614, 2008. ISSN 1093023X (ISSN). URL <http://ezproxy.deakin.edu.au/login?url=http://search.proquest.com/docview/62000189?accountid=10445%5Cnhttp://library.deakin.edu.au/resserv?genre=article{%&issn=1093023X{%&}title=Journal+of+Interactive+Learning+Research{%&}volume=19{%&}issue=4{%&}date=2008-10-01{%&}atitle=.> 188
- [166] J. A. Foster, P. K. Sheridan, R. Irish, and G. S. Frost. Gamification as a strategy for promoting deeper investigation in a reverse engineering activity. In *American Society for Engineering Education*. American Society for Engineering Education, 2012.
- [167] J. Fox and J. N. Bailenson. Virtual Self-Modeling: The Effects of Vicarious Reinforcement and Identification on Exercise Behaviors. *Media Psychology*, 12:1–25, 2009. ISSN 1521-3269. doi: 10.1080/15213260802669474. 137
- [168] P. Freire. *Pedagogy of the Oppressed*. 2000. ISBN 0826412769. URL <http://books.google.com/books?hl=en{%&}lr={%&}id=xFXFD414ioC{%&}oi=fnd{%&}pg=PA9{%&}dq=Pedagogy+of+the+Oppressed{%&}ots=sXT85aj3Yg{%&}sig=kGQxf5g5PayBMGM9HeBG1M1MK5c.> 36
- [169] B. H. Friedman. An autonomic flexibility–neurovisceral integration model of anxiety and cardiac vagal tone. *Biological Psychology*, 74(2):185–199, 2007. ISSN 03010511.

- doi: 10.1016/j.biopsycho.2005.08.009. URL <http://www.sciencedirect.com/science/article/pii/S0301051106001840>. 154
- [170] B. H. Friedman and J. F. Thayer. Anxiety and autonomic flexibility: A cardiovascular approach. *Biological Psychology*, 47(3):243–263, 1998. ISSN 03010511. doi: 10.1016/S0301-0511(97)00027-6. 154
- [171] R. S. Friedman and J. Förster. Implicit affective cues and attentional tuning: an integrative review. *Psychological bulletin*, 136(5):875–893, 2010. ISSN 0033-2909. doi: 10.1037/a0020495. 147
- [172] J. H. Frith. *Constructing location, one check-in at a time: examining the practices of Foursquare users*. North Carolina State University, 2012. 186
- [173] FromSoftware. FromSoftware, 2014. URL <http://www.fromsoftware.jp/>. 139
- [174] T. Fullerton. What games do well. *Postsecondary Play: The Role of Games and Social Media in Higher Education*, 2014. 167, 173
- [175] T. Fullerton. *Game design workshop: a playcentric approach to creating innovative games*. CRC press, 2014. 189
- [176] Gamepedia. Heroes, 2016. 39
- [177] C. M. Ganley, L. A. Mingle, A. M. Ryan, K. Ryan, M. Vasilyeva, and M. Perry. An examination of stereotype threat effects on girls' mathematics performance. *Developmental psychology*, 49(10):1886, 2013. 35
- [178] R. Garner, M. G. Gillingham, and C. S. White. Effects of "Seductive Details" on Macroprocessing and Microprocessing in Adults and Children. *Source: Cognition and Instruction*, 6(1):41–57, 1989. ISSN 0737-0008. doi: 10.1207/s1532690xci0601_2. 24, 26, 92, 94, 166, 167
- [179] R. Garris and R. Ahlers. A Game-Based Training Model: Development, Application,

- And Evaluation. In *The Interservice/Industry Training, Simulation & Education Conference (IITSEC)*, 2001. URL <http://ntsa.metapress.com/app/home/contribution.asp?referrer=parent{%&}backto=issue,6,151;journal,6,7;linkingpublicationresults,1:113340,1>. 173
- [180] R. Garris, R. Ahlers, and J. E. Driskell. Games, Motivation, and Learning: A Research and Practice Model. *Simulation & Gaming*, 33(4):441–467, 2002. ISSN 1046-8781. doi: 10.1177/1046878102238607. URL <http://sag.sagepub.com/content/33/4/441.short>. 167
- [181] M. Gåsland. Game mechanic based e-learning. *Science And Technology, Master Thesis (June 2011)*. Available at: <http://ntnu.diva-portal.org/smash/get/diva2,441760,2011>.
- [182] J. P. Gee. What video games have to teach us about learning and literacy. *Computers in Entertainment*, 1(1):20, oct 2003. ISSN 15443574. doi: 10.1145/950566.950595. 37
- [183] J. P. Gee. *What Video Games Have to Teach Us About Learning and Literacy*. Palgrave Macmillan, 2007. ISBN 1403984530. 31
- [184] J. P. Gee. Game-like learning: An example of situated learning and implications for opportunity to learn. *Assessment, Equity, and Opportunity to Learn*, pages 200–221, 2008. doi: <http://dx.doi.org/10.1017/CBO9780511802157.010>. 167, 172
- [185] A. Gekker. Health games: Taxonomy analysis and multiplayer design suggestions. *SGDA'12 Proceedings of the Third international conference on Serious Games Development and Applications*, 7528 LNCS:13–30, 2012. ISSN 03029743. doi: 10.1007/978-3-642-33687-4_2. 141
- [186] C. E. Gibson, J. Losee, and C. Vitiello. A replication attempt of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999): Identity salience and shifts in quantitative

- performance. *Social Psychology*, 45(3):194–198, 2014. ISSN 21512590. doi: 10.1027/1864-9335/a000184.
- [187] D. Gibson, N. Ostashewski, K. Flintoff, S. Grant, and E. Knight. Digital badges in education. *Education and Information Technologies*, (May 2016):1–8, 2013. ISSN 13602357. doi: 10.1007/s10639-013-9291-7. 186
- [188] B. G. Glaser. *Theoretical sensitivity: Advances in the methodology of grounded theory*. Sociology Pr, 1978.
- [189] I. Glover. Open Badges: A visual method of recognising achievement and increasing learner motivation. *Student Engagement and Experience Journal*, 2(1):1–4, 2013. ISSN 2047-9476. doi: 10.7190/seej.v1i1.66. URL <http://research.shu.ac.uk/SEEJ/index.php/seej/article/view/66>.
- [190] T. Gnambs, M. Appel, and B. Batinic. Color red in web-based knowledge testing. *Computers in Human Behavior*, 26(6):1625–1631, 2010. ISSN 07475632. doi: 10.1016/j.chb.2010.06.010. URL <http://linkinghub.elsevier.com/retrieve/pii/S0747563210001822>. 25, 93, 147, 148, 154, 209
- [191] G. Goehle. Gamification and web-based homework. *Primus*, 23(3):234–246, 2013. 186
- [192] E. T. Goetz and M. Sadoski. Commentary: The Perils of Seduction: Distracting Details or Incomprehensible Abstractions? *Reading Research Quarterly*, 30(3):500–511, 1995. ISSN 00340553. doi: 10.2307/747628. URL <http://www.jstor.org/stable/747628>. 173
- [193] M. E. Gonnerman Jr, C. P. Parker, H. Lavine, and J. Huff. The relationship between self-discrepancies and affective states: The moderating roles of self-monitoring and standpoints on the self. *Personality and Social Psychology Bulletin*, 26(7):810–819, 2000. 190
- [194] C. Good, J. Aronson, and M. Inzlicht. Improving adolescents’ standardized test

- performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, 24(6):645–662, 2003. ISSN 01933973. doi: 10.1016/j.appdev.2003.09.002. 34
- [195] C. Good, A. Rattan, and C. S. Dweck. Why do women opt out? Sense of belonging and women’s representation in mathematics. *Journal of personality and social psychology*, apr 2012. ISSN 1939-1315. doi: 10.1037/a0026659.
- [196] J. Goode and J. Margolis. *Exploring Computer Science*, 2011. ISSN 19466226.
- [197] A. Graesser, A. Witherspoon, B. McDaniel, S. D’Mello, P. Chipman, and B. Gholson. Detection of Emotions during Learning with AutoTutor. *Proceedings of the 28th Annual Meetings of the Cognitive Science Society*, 2006. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.2114{&rep=rep1{&}type=pdf>. 44
- [198] J. G. Greeno. The situativity of knowing, learning, and research. *American Psychologist*, 53(1):5–26, 1998. ISSN 0003-066X. doi: 10.1037/0003-066X.53.1.5. 167
- [199] J. G. Greeno, A. M. Collins, and L. B. Resnick. Cognition and learning. Handbook of educational psychology. In *Handbook of educational psychology*, pages 15–46. 1996. URL <http://psycnet.apa.org/psycinfo/1996-98614-001>.
- [200] M. Gresalfi, T. Martin, V. Hand, and J. Greeno. Constructing competence: An analysis of student participation in the activity systems of mathematics classrooms. *Educational Studies in Mathematics*, 70(1):49–70, 2009. ISSN 00131954. doi: 10.1007/s10649-008-9141-5.
- [201] S. Grover and R. Pea. Computational Thinking in K-12: A Review of the State of the Field. *Educational Researcher*, 42(1):38–43, 2013. ISSN 0013-189X. doi: 10.3102/0013189X12463051. URL <http://edr.sagepub.com/cgi/doi/10.3102/0013189X12463051>. 40, 41, 43, 44

- [202] R. E. Guadagno, J. Blascovich, J. N. Bailenson, and C. McCall. Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology*, 10(1): 1–22, 2007. ISSN 15213269. doi: 10.108/15213260701300865. [31](#), [130](#), [132](#), [136](#), [138](#), [184](#), [189](#)
- [203] J. Guegan, S. Buisine, F. Mantelet, N. Maranzana, and F. Segonds. Avatar-mediated creativity: When embodying inventors makes engineers more creative. *Computers in Human Behavior*, 61:165–175, 2016. ISSN 07475632. doi: 10.1016/j.chb.2016.03.024. [188](#)
- [204] A. Gulz and M. Haake. Design of animated pedagogical agents - A look at their look. *International Journal of Human Computer Studies*, 64(4):322–339, 2006. ISSN 10715819. doi: 10.1016/j.ijhcs.2005.08.006. [131](#), [138](#)
- [205] G. H. Guyatt, S. O. Pugsley, M. J. Sullivan, P. J. Thompson, L. Berman, N. L. Jones, E. L. Fallen, and D. W. Taylor. Effect of encouragement on walking test performance. *Thorax*, 39(11):818–822, 1984. ISSN 0040-6376. doi: 10.1136/thx.39.11.818. [155](#), [156](#), [164](#)
- [206] L. Hakulinen, T. Auvinen, and A. Korhonen. Empirical study on the effect of achievement badges in TRAKLA2 online learning environment. *Proceedings - 2013 Learning and Teaching in Computing and Engineering, LaTiCE 2013*, pages 47–54, 2013. ISSN 18630383. doi: 10.1109/LaTiCE.2013.34. [184](#), [186](#), [187](#)
- [207] M. Hall, H. National, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 2009. ISSN 19310145. doi: 10.1145/1656274.1656278. [102](#)
- [208] J. Hamari. Framework for Designing and Evaluating Game Achievements. *Proceedings of DiGRA 2011 Conference: Think Design Play*, page 20, 2011. ISSN ISSN 2342-9666. doi: 10.1.1.224.9966. URL <http://www.mendeley.com/catalog/framework-designing-evaluating-game-achievements/>. [187](#)

- [209] J. Hamari. Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service. *Electronic Commerce Research and Applications*, 12(4):236–245, 2013. ISSN 15674223. doi: 10.1016/j.elerap.2013.01.004. URL <http://dx.doi.org/10.1016/j.elerap.2013.01.004>. 184, 186, 187
- [210] J. Hamari. Do badges increase user activity? A field experiment on the effects of gamification. *Computers in Human Behavior*, 71:469–478, 2017. ISSN 07475632. doi: 10.1016/j.chb.2015.03.036. URL <http://dx.doi.org/10.1016/j.chb.2015.03.036>. 184, 186, 187
- [211] J. Hamari, J. Koivisto, and H. Sarsa. Does gamification work? - A literature review of empirical studies on gamification. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 3025–3034, 2014. ISSN 15301605. doi: 10.1109/HICSS.2014.377. 187
- [212] J. Hamilton. Identifying with an avatar: a multidisciplinary perspective. *Proceedings of the Cumulus Conference: 38th South: Hemispheric Shifts Across Learning, Teaching and Research*, (November):1–14, 2009. URL <http://eprints.qut.edu.au/29701/>. 32, 132, 190
- [213] T. Hansen, M. Olkkonen, S. Walter, and K. R. Gegenfurtner. Memory modulates color appearance. *Nature neuroscience*, 9(11):1367–8, 2006. ISSN 1097-6256. doi: 10.1038/nn1794. URL <http://www.ncbi.nlm.nih.gov/pubmed/17041591>. 147
- [214] M. D. Hanus and J. Fox. Persuasive avatars: The effects of customizing a virtual salesperson’s appearance on brand liking and purchase intentions. *International Journal of Human-Computer Studies*, 84:33–40, 2015.
- [215] P. L. Hardre and J. Reeve. A motivational model of rural students’ intentions to persist in, versus drop out of, high school. *Journal of Educational Psychology*, 95(2):347–

- 356, 2003. ISSN 1939-2176. doi: 10.1037/0022-0663.95.2.347. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-0663.95.2.347>. 168
- [216] J. Harms, D. Seitz, C. Wimmer, K. Kappel, and T. Grechenig. Low-Cost Gamification of Online Surveys : Improving the User Experience through Achievement Badges. *CHI PLAY'15 Proceedings of the 2nd ACM SIGCHI annual symposium on Computer-human interaction in play.*, pages 109–113, 2015. doi: 10.1145/2793107.2793146. URL <http://johannesharms.com/publications/harms2015badges.pdf>. 187
- [217] S. F. Harp and R. E. Mayer. How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology*, 90(3): 414–434, 1998. ISSN 0022-0663. doi: 10.1037/0022-0663.90.3.414. 167, 172
- [218] D. F. Harrell. Toward a Theory of Critical Computing: The Case of Social Identity Representation in Digital Media Applications. *CTheory*, 2010. 29, 31, 131, 132
- [219] D. F. Harrell. *Phantasmal Media: An Approach to Imagination, Computation, and Expression*. The MIT Press, 2013. ISBN 0262019337. 30
- [220] D. F. Harrell. Using Virtual Identities in Computer Science Learning for Broadening Participation. *NSF Annual Report*, 2017.
- [221] D. F. Harrell and S. V. Harrell. Imagination, Computation, and Self-Expression: Situated Character and Avatar Mediated Identity. *Leonardo electronic almanac*, 17(2): 74–91, jan 2012. ISSN 10714391. doi: 10.5900/SU_9781906897161_2012.17(2)_74. 39
- [222] D. F. Harrell, D. Kao, C. Lim, J. Lipshin, and A. Sutherland. The Chimeria Platform: User Empowerment through Expressing Social Group Membership Phenomena. *Digital Humanities*, 2014. 31
- [223] D. F. Harrell, S. Veeragoudar Harrell, D. Kao, D. Olson, A. Rodríguez, and L. Carney.

- Exploring the Use of Virtual Identities in Computer Science Learning for Broadening Participation (work in progress). 2017.
- [224] S. V. Harrell and D. Harrell. Exploring the Potential of Computational Self-Representations for Enabling Learning: Examining At-risk Youths' Development of Mathematical/Computational Agency. *Proceedings of the Digital Arts and Culture Conference*, 2009. [132](#), [137](#), [172](#)
- [225] C. Harteveld and S. Sutherland. The Goal of Scoring: Exploring the Role of Game Performance in Educational Games. *Proceedings of the 33rd annual ACM conference on Human factors in computing systems (CHI 2015)*, 2015. [20](#), [130](#), [131](#), [136](#), [140](#), [168](#), [172](#)
- [226] C. Harteveld, G. Smith, G. Carmichael, E. Gee, and C. Stewart-Gardiner. A Design-Focused Analysis of Games Teaching Computer Science Methodology for Examining Games that Teach Computer Science. 2014. [45](#)
- [227] C. Hecker. Achievements considered harmful? *GDC*, 2010. URL <http://www.gdcvault.com/play/1012970/Achievements-Considered-Harmful>.
- [228] D. Hefner, C. Klimmt, and P. Vorderer. Identification with the Player Character as Determinant of Video Game Enjoyment. In *Entertainment Computing - ICEC 2007*, pages 39–48. 2007. ISBN 978-3-540-74872-4, 978-3-540-74873-1 SV - 4740. doi: 10.1007/978-3-540-74873-1_6. URL http://link.springer.com.libproxy.usc.edu/chapter/10.1007/978-3-540-74873-1_{_}6_{%}5Cnfiles/12188/10.1007_{_}978-3-540-74873-1_{_}6.pdf_{%}5Cnfiles/11935/978-3-540-74873-1_{_}6.html. [184](#), [189](#)
- [229] J. Henderlong and M. R. Lepper. The effects of praise on children's intrinsic motivation: a review and synthesis. *Psychological bulletin*, 128(5):774–795, 2002. ISSN 0033-2909. doi: 10.1037/0033-2909.128.5.774. [164](#)

- [230] P. Hernandez. The Most Popular Levels In Super Mario Maker (So Far), 2015. URL <http://kotaku.com/the-most-popular-levels-in-super-mario-maker-so-far-1730170457>. 209
- [231] J. C. Herz. Harnessing the hive. *Creative industries*, pages 327–341, 2005. 39
- [232] T. M. Hess, C. Auman, S. J. Colcombe, and T. A. Rahhal. The impact of stereotype threat on age differences in memory performance. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 58(1):P3—P11, 2003. 33
- [233] A. Hicks, B. Peddycord III, and T. Barnes. Building games to learn from their players: Generating hints in a serious game. In *International Conference on Intelligent Tutoring Systems*, pages 312–317. Springer, 2014. 45
- [234] E. T. Higgins. Self-discrepancy: a theory relating self and affect. *Psychological review*, 94(3):319, 1987. 189, 190, 207
- [235] R. a. Hill and R. a. Barton. Psychology: red enhances human performance in contests. *Nature*, 435(May):293, 2005. ISSN 0028-0836. doi: 10.1038/435293a. 147, 148, 152
- [236] D. C. Hoaglin and B. Iglewicz. Fine-Tuning Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association*, 82(400): 1147–1149, 1987. ISSN 0162-1459. doi: 10.1080/01621459.1987.10478551. URL <http://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478551>. 159
- [237] C. Hoffner. Children’s wishful identification and parasocial interaction with favorite television characters. *Journal of Broadcasting & Electronic Media*, 40(3):389–402, 1996. 189, 190, 207
- [238] C. Hoffner and M. Buchanan. Young adults’ wishful identification with television characters: The role of perceived similarity and character attributes. *Media psychology*, 7(4):325–351, 2005. 189, 190, 207

- [239] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 1993. ISSN 08856125. doi: 10.1023/A:1022631118932. 103
- [240] R. Hooi and H. Cho. Deception in avatar-mediated virtual environment. *Computers in Human Behavior*, 29(1):276–284, 2013. 189
- [241] Y. Hori, Y. Tokuda, T. Miura, A. Hiyama, and M. Hirose. Communication pedometer: a discussion of gamified communication focused on frequency of smiles. In *Proceedings of the 4th augmented human international conference*, pages 206–212. ACM, 2013. 186
- [242] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425, 2011. 80
- [243] C. Huff, D. Johnson, and K. Miller. Virtual harms and real responsibility. *IEEE Technology and Society Magazine*, 22(2):12–19, 2003. ISSN 0278-0097. doi: 10.1109/MTAS.2003.1216238. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1216238>. 31, 131, 132, 136
- [244] B. Hulshof. The influence of colour and scent on people’s mood and cognitive performance in meeting rooms. *Master Thesis*, (May):1–97, 2013. 25, 93, 147, 148, 154, 193
- [245] A. C. Hurlbert and Y. Ling. Biological components of sex differences in color preference. *Current Biology*, 17(16):623–625, 2007. ISSN 09609822. doi: 10.1016/j.cub.2007.06.022. 152
- [246] J. Hutchings. Folklore and Symbolism of Green. *Folklore*, 108:55–63, 1997. ISSN 0015-587X. doi: 10.1080/0015587X.1997.9715937. URL <http://www.jstor.org/stable/1260708>{%}5Cn<http://www.jstor.org/stable/pdfplus/1260708.pdf?acceptTC=true>. 147

- [247] E. Hutchins. Distributed Cognition, 2000. ISSN 14355558. URL <http://www.slis.indiana.edu/faculty/yrogers/dist{ }cog/>.
- [248] A. E. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3D avatar creation from handheld video input. *ACM Transactions on Graphics*, 34(4):45:1–45:14, 2015. ISSN 07300301. doi: 10.1145/2766974. URL <http://dl.acm.org/citation.cfm?doid=2809654.2766974>. 39
- [249] W. IJsselsteijn, Y. De Kort, K. Poels, A. Jurgelionis, and F. Bellotti. Characterising and Measuring User Experiences in Digital Games. *International Conference on Advances in Computer Entertainment Technology*, 620:1–4, 2007. doi: 10.1007/978-1-60761-580-4. URL <http://repository.tue.nl/661449>. 78, 123, 124, 134, 143, 147, 149, 155, 158, 169
- [250] W. IJsselsteijn, K. Poels, and Y. A. W. de Kort. The Game Experience Questionnaire: Development of a self-report measure to assess player experiences of digital games. *TU Eindhoven, Eindhoven, The Netherlands*, 2008. 78
- [251] A. Ilie, S. Ioan, L. Zagrean, and M. Moldovan. Better to be red than blue in virtual competition. *Cyber Psychology & Behavior*, 11(3):375–377, 2008. ISSN 1094-9313. doi: 10.1089/cpb.2007.0122. 25, 93, 147, 148, 154
- [252] M. Inzlicht and T. Ben-Zeev. A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, 11(5):365–371, 2000. 34
- [253] M. Inzlicht, L. McKay, and J. Aronson. Stigma as ego depletion: How being the target of prejudice affects self-control. *Psychological Science*, 17(3):262–269, 2006. 34
- [254] S. Ioan, M. Sandulache, S. Avramescu, A. Ilie, A. Neacsu, L. Zagrean, and M. Moldovan. Red is a distractor for men in competition. *Evolution and Human*

- Behavior*, 28(4):285–293, 2007. ISSN 10905138. doi: 10.1016/j.evolhumbehav.2007.03.001. 147, 148
- [255] P. Ipeiritis. The New Demographics of Mechanical Turk. URL <http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html>. 80
- [256] K. Isbister and C. Nass. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, 53:251–267, 2000. ISSN 10715819. doi: 10.1006/ijhc.2000.0368. URL <http://www.sciencedirect.com/science/article/pii/S1071581900903689>. 31, 132, 136, 189
- [257] M. Ito. *Engineering play: A cultural history of children's software*. MIT Press, 2012. 38, 184
- [258] G. H. Jacobs. *Comparative Color Vision*. 1981. ISBN 9780123785206. doi: 10.1016/B978-0-12-378520-6.50010-8. URL <http://www.sciencedirect.com/science/article/pii/B9780123785206500108>. 147
- [259] M. Jakobsson. The achievement machine: Understanding Xbox 360 achievements in gaming practices. *Game Studies*, 11(1), 2011. ISSN 16047982. 184, 186, 187
- [260] R. Jarman. Science learning through scouting: an understudied context for informal science education. *International Journal of Science Education*, 27(4):427–450, 2005. 186
- [261] F. Jiang, S. Lu, X. Yao, X. Yue, and W. tung Au. Up or down? How culture and color affect judgments. *Journal of Behavioral Decision Making*, 27(3):226–234, 2014. ISSN 10990771. doi: 10.1002/bdm.1800. 148
- [262] S.-A. A. Jin. "I Feel More Connected to the Physically Ideal Mini Me than the Mirror-Image Mini Me": Theoretical Implications of the "Malleable Self" for Speculations

- on the Effects of Avatar Creation on Avatar–Self Connection in Wii. *Cyberpsychology, Behavior, and Social Networking*, 13(5):567–570, 2010. 188
- [263] O. P. John and S. Srivastava. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(510):102–138, 1999. ISSN 00031224. doi: citeulike-article-id:3488537. URL <http://books.google.com/books?hl=en&lr=&id=b0yalwi1HDMC&oi=fnd&pg=PA102&dq=The+big-five+trait+taxonomy:+History,+Measurement,+and+Theoretical+Perspectives.&ots=756zS6ZtPk&sig=-3pfI7eNKlyZLlJYEmwdDYeJ82Y{%}5Cnhttp://scholar.google.de/scholar?hl=de&q=john+sriva>. 164
- [264] M. Johns, M. Inzlicht, and T. Schmader. Stereotype threat and executive resource depletion: examining the influence of emotion regulation. *Journal of Experimental Psychology: General*, 137(4):691, 2008. 34
- [265] A. M. Johnson, M. D. Didonato, and M. Reisslein. Animated agents in K-12 engineering outreach: Preferred agent characteristics across age levels. *Computers in Human Behavior*, 29(4):1807–1815, 2013. ISSN 07475632. 31, 132, 136
- [266] K. Jolivette, J. H. Wehby, J. Canale, and N. G. Massey. Effect of choice-making opportunities on the behavior of students with emotional and behavioral disorders. *Behavioral Disorders*, 26(2):131–145, 2001. ISSN 0198-7429, 0198-7429. URL http://proxy2.lib.umanitoba.ca/login?url=http://search.proquest.com/docview/620004188?accountid=14569{%}5Cnhttp://sfxhosted.exlibrisgroup.com/umanitoba?url{__}ver=Z39.88-2004&rft{__}val{__}fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ:psycinfo&atitle=. 168, 197
- [267] B. M. Jones. Individualistic vs. Competitive Participation: The Effect on Intrinsic

- Motivation. 1985. 165
- [268] E. Joosten, G. Lankveld, and P. Spronck. Colors and emotions in video games. *11th International Conference on Entertainment Computing*, (Figure 1), 2010. 148, 153
- [269] B. Joseph. Six ways to look at badging systems designed for learning. *Global Kids: Online Leadership Program*, [online] Available at: <http://www.olpglobalkids.org/content/six-ways-look-badging-systems-designed-learning>, 2012. 186
- [270] J. Jovanovic and V. Devedzic. Open Badges: Novel Means to Motivate, Scaffold and Recognize Learning. *Technology, Knowledge and Learning*, 20(1):115–122, 2015. ISSN 22111662. doi: 10.1007/s10758-014-9232-6.
- [271] I. Jung, M. Kim, and K. Han. Red for Romance, Blue for Memory. In *HCI International 2011 - Posters Extended Abstracts*, volume 173, pages 284–288. 2011. ISBN 978-3-642-22097-5. doi: 10.1007/978-3-642-22098-2_57. URL http://dx.doi.org/10.1007/978-3-642-22098-2_{_}57. 25, 93, 147, 148, 154
- [272] L. Jusim. Is Stereotype Threat Overcooked, Overstated, and Oversold?, 2016. URL <https://www.psychologytoday.com/blog/rabble-rouser/201512/is-stereotype-threat-overcooked-overstated-and-oversold>. 35
- [273] Y. B. Kafai and Q. Burke. Constructionist Gaming: Understanding the Benefits of Making Games for Learning. *Educational Psychologist*, 2015. 208
- [274] Y. B. Kafai and Q. Burke. Constructionist Gaming: Understanding the Benefits of Making Games for Learning. *Educational Psychologist*, 50(4):313–334, 2015. ISSN 0046-1520. doi: 10.1080/00461520.2015.1124022. URL <http://www.tandfonline.com/doi/full/10.1080/00461520.2015.1124022>. 190
- [275] Y. B. Kafai, D. A. Fields, and M. S. Cook. Your Second Selves: Player-

- Designed Avatars. *Games and Culture*, 5(1):23–42, 2010. ISSN 1555-4120. doi: 10.1177/1555412009351260. URL <http://gac.sagepub.com/cgi/doi/10.1177/1555412009351260>. 39
- [276] D. Kao. MazeStar: A Platform for Studying Virtual Identity and Computer Science Education. In *Foundations of Digital Games ASETGC Workshop 2017*, 2017. ISBN 9781450353199. 190
- [277] D. Kao and D. F. Harrell. Exploring the Impact of Role Model Avatars on Game Experience in Educational Games. *The ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHI PLAY)*, 2015. 22, 24, 35, 92, 188, 208, 213
- [278] D. Kao and D. F. Harrell. Mazzy: A STEM Learning Game. *Foundations of Digital Games*, 2015. 47, 124, 133, 142, 149, 156, 168, 190
- [279] D. Kao and D. F. Harrell. Exploring the Use of Role Model Avatars in Educational Games. In *Proceedings of the AIIDE Workshop on Experimental AI in Games, co-located with Artificial Intelligence in Interactive Digital Entertainment*, 2015. 22, 24, 35, 92, 188, 211, 213
- [280] D. Kao and D. F. Harrell. Toward Avatar Models to Enhance Performance and Engagement in Educational Games. In *Computational Intelligence in Games*, 2015. 132, 137
- [281] D. Kao and D. F. Harrell. Exploring the Impact of Role Model Avatars on Game Experience in Educational Games. In *The ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHI PLAY)*, 2015. 211
- [282] D. Kao and D. F. Harrell. Toward Avatar Models to Enhance Performance and Engagement in Educational Games. *IEEE Computational Intelligence in Games*, 2015. 131, 211
- [283] D. Kao and D. F. Harrell. Toward Evaluating the Impacts of Virtual Identities on STEM Learning. *Foundations of Digital Games*, 2015. 19, 32, 108, 113

- [284] D. Kao and D. F. Harrell. Exigent: An Automatic Avatar Generation System. *10th International Conference on the Foundations of Digital Games*, 2015. [40](#), [71](#)
- [285] D. Kao and D. F. Harrell. Toward Understanding the Impacts of Role Model Avatars on Engagement in Computer Science Learning. In *The annual meeting of the American Educational Research Association (AERA)*, 2016. [22](#), [24](#), [35](#), [92](#), [132](#), [188](#), [211](#), [213](#)
- [286] D. Kao and D. F. Harrell. Exploring the Effects of Dynamic Avatars on Performance and Engagement in Educational Games. In *Games+Learning+Society (GLS 2016)*, 2016. [22](#), [24](#), [25](#), [92](#), [93](#), [211](#), [213](#)
- [287] D. Kao and D. F. Harrell. Exploring the Impact of Avatar Color on Game Experience in Educational Games. *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI 2016)*, 2016. [25](#), [93](#), [193](#), [211](#), [213](#)
- [288] D. Kao and D. F. Harrell. Exploring the Effects of Encouragement in Educational Games. *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI 2016)*, 2016. [22](#), [26](#), [94](#), [211](#), [213](#)
- [289] D. Kao and D. F. Harrell. Toward Understanding the Impact of Visual Themes and Embellishment on Performance, Engagement, and Self-Efficacy in Educational Games. *The annual meeting of the American Educational Research Association (AERA)*, 2017. [19](#), [22](#), [26](#), [94](#), [212](#), [214](#)
- [290] D. Kao and D. F. Harrell. The Effects of Badges and Avatar Identification on Play and Making in Educational Games. In *CHI*, 2018. [22](#), [23](#), [25](#), [93](#), [212](#), [213](#)
- [291] D. Kao, D. F. Harrell, C.-U. Lim, S. V. Harrell, M. Wagoner, and H. Ho. Highlighting MazeStar: A Platform for Studying Avatar Use in Computer Science Learning Environments. *Games+Learning+Society (GLS 2016)*, 2016. [20](#), [39](#)
- [292] I. Katz and A. Assor. When choice motivates and when it does not. *Educational*

- Psychology Review*, 19(4):429–442, 2007. ISSN 1040726X. doi: 10.1007/s10648-006-9027-y. [171](#), [207](#)
- [293] G. Kaufman, M. Flanagan, and M. Seidman. Creating Stealth Game Interventions for Attitude and Behavior Change : An Embedded Design Model. *DiGRA 2015: Diversity of Play*, 2015. [132](#)
- [294] C. Kelleher, R. Pausch, and S. Kiesler. Storytelling alice motivates middle school girls to learn computer programming. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, page 1455, 2007. ISSN 00219517. [45](#)
- [295] M. Kernan, B. Heimann, and P. Hanges. Effects of goal choice, strategy choice, and feedback source on goal acceptance, performance, and subsequent goals. *Journal of Applied Social Psychology*, 21:713–733, 1991. [168](#)
- [296] S. a. Khan, W. J. Levine, S. D. Dobson, and J. D. Kralik. Red signals dominance in male rhesus macaques. *Psychological science : a journal of the American Psychological Society / APS*, 22(8):1001–1003, 2011. ISSN 0956-7976. doi: 10.1177/0956797611415543. [147](#)
- [297] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, and Others. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011. [186](#), [208](#), [214](#)
- [298] C. Kim, S. G. Lee, and M. Kang. I became an attractive person in the virtual world: Users' identification with virtual communities and avatars. *Computers in Human Behavior*, 28(5):1663–1669, 2012. ISSN 07475632. doi: 10.1016/j.chb.2012.04.004. URL <http://dx.doi.org/10.1016/j.chb.2012.04.004>. [184](#), [189](#)
- [299] Y. Kim and A. L. Baylor. Pedagogical agents as learning companions: The role of agent competency and type of interaction. *Educational Technology Research and*

- Development*, 54(3):223–243, 2006. ISSN 10421629. [31](#), [130](#), [132](#), [136](#), [138](#), [184](#), [189](#)
- [300] Y. Kim and S. S. Sundar. Visualizing ideal self vs. actual self through avatars: Impact on preventive health outcomes. *Computers in Human Behavior*, 28(4):1356–1364, 2012. [189](#)
- [301] C. Kinzer, D. Hoffman, and S. Turkay. The impact of choice and feedback on learning, motivation, and performance in an educational video game. *GLS*, 2:175–181, 2012. URL http://gamesresearchlab.com/wp-content/uploads/2015/11/The_{ }Impact_{ }of_{ }Choice_{ }and_{ }Feedback.pdf. [197](#)
- [302] C. Klimmt. Dimensions and determinants of the enjoyment of playing digital games: A three-level model. In *Level up: Digital games research conference*, pages 246–257, 2003. [189](#)
- [303] A. N. Kluger and A. DeNisi. The effects of feedback interventions on performance: a historical review, a meta-analysis and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2):254–284, 1996. ISSN 0033-2909. doi: 10.1037/0033-2909.119.2.254. [26](#), [94](#), [155](#), [164](#)
- [304] D. Knuth. Literate programming. *The Computer Journal*, 1984. [19](#)
- [305] S. Koch and M. Bierbamer. Opening your product: impact of user innovations and their distribution platform on video game success. *Electronic Markets*, pages 1–12, 2016. [39](#)
- [306] A. M. Koenig and A. H. Eagly. Stereotype threat in men on a test of social sensitivity. *Sex Roles*, 52(7-8):489–496, 2005. [33](#)
- [307] K. Koenig and M. Hanson. Fueling interest in science: An after-school program model that works. *Science Scope*, 32(4):48–51, 2008.
- [308] K. H. Koh, A. Basawapatna, V. Bennett, and A. Repenning. Towards the Au-

- Automatic Recognition of Computational Thinking for Adaptive Visual Language Learning. *Visual Languages ...*, pages 59–66, 2010. ISSN 978-1-4244-7621-3. doi: 10.1109/VLHCC.2010.17. URL http://ieeexplore.ieee.org/xpls/abs/_all.jsp?arnumber=5635189. 44, 45
- [309] M. Kölling. The greenfoot programming environment. *ACM Transactions on Computing Education (TOCE)*, 10(4):14, 2010. 45
- [310] J. L. Kolodner, P. J. Camp, D. Crismond, B. Fasse, J. Gray, J. Holbrook, S. Puntambekar, and M. Ryan. Problem-Based Learning Meets Case-Based Reasoning in the Middle School Science Classroom: Putting Learning by Design Into Practice. *Journal of the Learning Sciences*, 12(4):495–547, 2009. ISSN 1050-8406. doi: 10.1207/S15327809JLS1204. 3
- [311] E. A. Konijn, M. Nije Bijvank, and B. J. Bushman. I wish I were a warrior: the role of wishful identification in the effects of violent video games on aggression in adolescent boys. *Developmental psychology*, 43(4):1038, 2007. 189
- [312] B. Krenn, S. Würth, and A. Hergovich. The impact of feedback on goal setting and task performance: Testing the feedback intervention theory. *Swiss Journal of Psychology*, 2013. 164
- [313] Á. Kristjánsson, P. Vuilleumier, P. Malhotra, M. Husain, and J. Driver. Priming of color and position during visual search in unilateral spatial neglect. *Journal of cognitive neuroscience*, 17:859–873, 2005. ISSN 0898-929X. doi: 10.1162/0898929054021148. 147
- [314] C. Kuhbandner and R. Pekrun. Joint effects of emotion and color on memory. *Emotion (Washington, D.C.)*, 13(3):375–9, 2013. ISSN 1931-1516. doi: 10.1037/a0031821. URL <http://www.ncbi.nlm.nih.gov/pubmed/23527500>. 25, 93, 147, 148, 154
- [315] M. Lakoff, George; Johnson. *Metaphors we live by*. 2003. ISBN 0226468011. 31

[316] P. J. Lang. The Emotion Probe. *American Psychologist Association*, 50(5):372–385, 1995. ISSN 0003-066X. doi: 10.1037/0003-066X.50.5.372. 147

[317] J. Lave. Situating learning in communities of practice. *Perspectives on socially shared cognition*, pages 63–82, 1991. ISSN 18780369. doi: 10.1037/10096-003.

URL [http://pitt.summon.serialssolutions.com/2.0.0/](http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwxV1NS8NAEB20BREPWrXxE{ }YPJE2abD5Ogh-)

[link/0/eLvHCXMwxV1NS8NAEB20BREPWrXxE{ }](http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwxV1NS8NAEB20BREPWrXxE{ }YPJE2abD5Ogh-)

[loDcLeiub3WwptKml7cGj{ }9yZTTZg1bNQKEsuZWfzZmb73huAsO{ }57hYmYJci40I](http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwxV1NS8NAEB20BREPWrXxE{ }YPJE2abD5Ogh-)

[bbUjNX5hSdc-SpueZZz1iINJuWJsSl](http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwxV1NS8NAEB20BREPWrXxE{ }YPJE2abD5Ogh-). 167

[318] J. Lave and E. Wenger. Situated learning: Legitimate peripheral participation. *Learning in doing*, 95:138, 1991. ISSN 00027294. doi: 10.2307/2804509. URL <http://books.google.com/books?id=CAVIOrW3vYAC{&}pgis=1>. 37, 167

[319] M. J. Lee, A. J. Ko, and I. Kwan. In-game assessments increase novice programmers’ engagement and level completion speed. *Proceedings of the ninth annual international ACM conference on International computing education research - ICER '13*, page 153, 2013. doi: 10.1145/2493394.2493410. URL <http://dl.acm.org/citation.cfm?doid=2493394.2493410>. 45

[320] O. Lee and A. Luykx. *Science education and student diversity: Synthesis and research agenda*. Cambridge University Press, 2006. 188

[321] S. Lehman, G. Schraw, M. T. McCrudden, and K. Hartley. Processing and recall of seductive details in scientific text. *Contemporary Educational Psychology*, 32(4): 569–587, 2007. ISSN 0361476X. doi: 10.1016/j.cedpsych.2006.07.002. 167, 172

[322] J. Lemke. Cognition, context, and learning: A social semiotic perspective. *Situated Cognition: Social, Semiotic, and Psychological Perspectives*, (1972):37–55, 1997. URL <http://books.google.com/books?id=ivqlQw9Dy-8C{&}pgis=1>. 167

[323] J. C. Lester, S. a. Converse, S. E. Kahler, S. T. Barlow, B. a. Stone, and R. S. Bhogal.

- The Persona Effect: Affective Impact of Animated Pedagogical Agents. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI '97*, pages 359–366, 1997. doi: 10.1145/258549.258797. 130, 138
- [324] Z. Lewis, M. C. Swartz, and E. Lyons. What's the Point?: A Review of Reward Systems Implemented in Gamification Interventions. *Games for Health Journal*, 5(2):93–99, 2016. ISSN 2161-7856. doi: 10.1089/g4h.2015.0078. 184, 187
- [325] B. J. Li and M. O. Lwin. Player see, player do: Testing an exergame motivation model based on the influence of the self avatar. *Computers in Human Behavior*, 59:350–357, 2016. ISSN 07475632. doi: 10.1016/j.chb.2016.02.034. URL <http://dx.doi.org/10.1016/j.chb.2016.02.034>. 189
- [326] D. D. Li, A. K. Liao, and A. Khoo. Player–Avatar Identification in video gaming: Concept and measurement. *Computers in Human Behavior*, 29(1):257–263, 2013. 188
- [327] W. Li, T. Grossman, and G. Fitzmaurice. GamiCAD: a gamified tutorial system for first time autocad users. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 103–112. ACM, 2012.
- [328] Z. Li, K.-W. Huang, and H. Cavusoglu. Quantifying the impact of badges on user engagement in online Q&A communities. 2012. 184, 187, 207
- [329] S. Lichtenfeld, M. a. Maier, A. J. Elliot, and R. Pekrun. The semantic red effect: Processing the word red undermines intellectual performance. *Journal of Experimental Social Psychology*, 45(6):1273–1276, 2009. ISSN 00221031. doi: 10.1016/j.jesp.2009.06.003. URL <http://dx.doi.org/10.1016/j.jesp.2009.06.003>. 25, 93, 147, 148, 154
- [330] C. Lim and D. Harrell. Developing Social Identity Models of Players from Game Telemetry Data. *AIIDE*, 2014. 31
- [331] C.-U. Lim and D. F. Harrell. Modeling Player Preferences in Avatar Customization

- Using Social Network Data. In *Computational Intelligence in Games*, 2013. ISBN 9781467353113. 31
- [332] C.-U. Lim and D. F. Harrell. Toward Telemetry-driven Analytics for Understanding Players and their Avatars in Videogames. *CHI Extended Abstracts*, pages 1175–1180, 2015. 72
- [333] C.-U. Lim and D. F. Harrell. Understanding Players’ Identities and Behavioral Archetypes from Avatar Customization Data. *CIG*, 2015. 72
- [334] S. Lim and B. Reeves. Being in the Game: Effects of Avatar Choice and Point of View on Psychophysiological Responses During Play. *Media Psychology*, 12(4):348–370, 2009. ISSN 1521-3269. doi: 10.1080/15213260903287242. URL <http://www.tandfonline.com/doi/pdf/10.1080/15213260903287242>. 189, 199, 201
- [335] R. J. Lin and X. Zhu. Leveraging social media for preventive care-A gamification system and insights. *Studies in health technology and informatics*, 180:838–842, 2012. 187
- [336] J. M. Linebarger and G. D. Kessler. The effect of avatar connectedness on task performance. *Lehigh Univ TR*, 2002. 36
- [337] J. Ling and M. Blades. Further Evidence for Automatic Encoding of Colour by Children and Adults. *British Journal of Developmental Psychology*, 20(4):537–544, 2002. ISSN 2044-835X. doi: 10.1348/026151002760390936. URL <https://login.iris.etsu.edu:3443/login?url=http://search.proquest.com/docview/85570605?accountid=10771>. 147
- [338] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, and Others. Galaxy Zoo: morpholo-

- gies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 2008. 186
- [339] I. J. Livingston, C. Gutwin, R. L. Mandryk, and M. Birk. How players value their characters in world of warcraft. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, pages 1333–1343, 2014. doi: 10.1145/2531602.2531661. URL <http://dl.acm.org/citation.cfm?doid=2531602.2531661>. 209
- [340] P. Lockwood. "Someone like me can be successful": Do college students need same-gender role models? *Psychology of Women Quarterly*, 30(1):36–46, mar 2006. ISSN 03616843. doi: 10.1111/j.1471-6402.2006.00260.x. URL <http://pwq.sagepub.com/content/30/1/36>. 34, 35, 129, 188, 192, 209
- [341] P. Lockwood and Z. Kunda. Superstars and me: Predicting the impact of role models on the self. *Journal of Personality and Social Psychology*, 73(1):91–103, 1997. ISSN 0022-3514. doi: 10.1037//0022-3514.73.1.91. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.73.1.91><http://psycnet.apa.org/journals/psp/73/1/91/>. 34, 123, 208
- [342] J. Logan. *Scents and the City*, 2009.
- [343] Logo Foundation. Logo (<http://el.media.mit.edu/logo-foundation/>), 2017. 45
- [344] S. C. Lonial and S. Van Auken. Wishful identification with fictional characters: An assessment of the implications of gender in message dissemination to children. *Journal of Advertising*, 15(4):4–42, 1986. 207
- [345] MacArthur Foundation. *Better Futures for 2 Million Americans Through Open Badges*, 2013. 186
- [346] E. E. Maccoby and W. C. Wilson. Identification and observational learning from films.

- The Journal of abnormal and social psychology*, 55(1):76, 1957. ISSN 0096-851X. doi: 10.1037/h0043015. [32](#), [132](#), [190](#)
- [347] P. Maes and Others. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, 1994. [27](#)
- [348] T. Mahlmann, A. Drachen, J. Togelius, A. Canossa, and G. N. Yannakakis. Predicting player behavior in Tomb Raider: Underworld. *Computational Intelligence in Games*, 2010. doi: 10.1109/ITW.2010.5593355. [102](#)
- [349] T. W. Malone and M. R. Lepper. Making learning fun: A taxonomy of intrinsic motivations for learning, 1987. ISSN 00376337. URL <http://hal.www.mendeley.com/research/making-learning-fun-a-taxonomy-of-intrinsic-motivations-for-learning/>. [167](#)
- [350] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann. Design lessons from the fastest q&a site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 2857–2866. ACM, 2011. [186](#)
- [351] A. Marder. Stack overflow badges and user behavior: An econometric approach. *IEEE International Working Conference on Mining Software Repositories*, 2015-Augus:450–453, 2015. ISSN 21601860. doi: 10.1109/MSR.2015.61. [184](#), [186](#), [187](#)
- [352] J. Marder and M. Fritz. The Internet’s hidden science factory. URL <http://www.pbs.org/newshour/updates/inside-amazons-hidden-science-factory/>. [80](#)
- [353] F. Martin and M. Resnick. Lego/Logo and electronic bricks: Creating a scienceland for children. In *Advanced educational technologies for mathematics and science*, pages 61–89. Springer, 1993. [38](#)

- [354] F. G. Martin. *Circuits to control–learning engineering by designing LEGO robots*. PhD thesis, Massachusetts Institute of Technology, 1994. [38](#)
- [355] J. E. Martin, D. E. Mithaug, P. Cox, L. Y. Peterson, J. L. Van Dycke, and M. E. Cash. Increasing self-determination: Teaching students to plan, work, evaluate, and adjust. *Exceptional Children*, 69(4):431–447, 2003. ISSN 0014-4029, 0014-4029. doi: doi:10.1177/001440290306900403. [168](#), [197](#)
- [356] D. M. Marx and P. A. Goff. Clearing the air: the effect of experimenter race on target’s test performance and subjective experience. *The British journal of social psychology / the British Psychological Society*, 44:645–657, 2005. ISSN 0144-6665. doi: 10.1348/014466604X17948. [113](#), [122](#), [129](#)
- [357] D. M. Marx and J. S. Roman. Female Role Models: Protecting Women’s Math Test Performance. *Personality and Social Psychology Bulletin*, 28:1183–1193, 2002. ISSN 0146-1672. doi: 10.1177/01461672022812004. [34](#), [35](#), [113](#), [129](#), [184](#), [188](#), [192](#), [209](#)
- [358] D. M. Marx, D. a. Stapel, and D. Muller. We can do it: the interplay of construal orientation and social comparisons under threat. *Journal of personality and social psychology*, 88(3):432–446, 2005. ISSN 0022-3514. doi: 10.1037/0022-3514.88.3.432. [34](#), [129](#), [188](#), [192](#), [209](#)
- [359] D. M. Marx, S. J. Ko, and R. a. Friedman. The “Obama Effect”: How a salient role model reduces race-based performance differences. *Journal of Experimental Social Psychology*, 45(4):953–956, jul 2009. ISSN 00221031. doi: 10.1016/j.jesp.2009.03.012. URL <http://linkinghub.elsevier.com/retrieve/pii/S0022103109000742>. [129](#)
- [360] W. Mason and S. Suri. Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, 2012. ISSN 1554-3528. doi: 10.3758/s13428-011-0124-6. URL <http://www.springerlink.com/index/10.3758/s13428-011-0124-6>. [21](#), [80](#)

- [361] M. Mateas. Procedural literacy: educating the new media practitioner. *On the Horizon*, 13(2):101–111, 2005. ISSN 1074-8121. doi: 10.1108/10748120510608133. URL <http://www.emeraldinsight.com/10.1108/10748120510608133>. 40
- [362] Math is Fun - Maths Resources. Shapes, 2014. URL <http://www.mathsisfun.com/shape.html>. 99
- [363] R. E. Mayer. *The Cambridge handbook of multimedia learning*, volume 16. 2005. ISBN 9780521838733. doi: 10.1075/idj.16.1.13pel. 167
- [364] R. E. Mayer, J. Heiser, and S. Lonn. Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology*, 93(1):187–198, 2001. ISSN 1939-2176. doi: 10.1037/0022-0663.93.1.187. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-0663.93.1.187>. 173
- [365] R. E. Mayer, M. Hegarty, S. Mayer, and J. Campbell. When static media promote active learning: annotated illustrations versus narrated animations in multimedia instruction. *Journal of experimental psychology. Applied*, 11(4):256–265, 2005. ISSN 1076-898X. doi: 10.1037/1076-898X.11.4.256. 173
- [366] A. Mazur, A. Booth, and J. M. D. Jr. Testosterone and Chess Competition. *Social Psychology Quarterly*, 55(1):70, 1992. ISSN 01902725. doi: 10.2307/2786687. 165
- [367] E. McAuley, T. Duncan, and V. V. Tammen. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport*, 60(1):48–58, 1989. 79, 198
- [368] J. Mccoy, M. Treanor, B. Samuel, B. Tearse, M. Mateas, and N. Wardrip-fruin. Comme il Faut. *Journal of Affective Disorders*, 40(1-2):3, 2010. doi: 10.1016/0165-0327(96)82444-X. URL <http://dl.acm.org/citation.cfm?id=1822309.1822319>. 154

- [369] R. McDaniel, R. Lindgren, and J. Friskics. Using badges for shaping interactions in online learning environments. In *Professional Communication Conference (IPCC), 2012 IEEE International*, pages 1–4. IEEE, 2012. [186](#)
- [370] D. G. McDonald and H. Kim. When I die, I feel small: Electronic game characters and the social self. *Journal of Broadcasting & Electronic Media*, 45(2):241–258, 2001. [189](#)
- [371] McGraw-Hill Education. SmartBook. 2015. [44](#)
- [372] R. B. McIntyre, R. M. Paulson, and C. G. Lord. Alleviating women’s mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology*, 39(1):83–90, 2003. ISSN 00221031. doi: 10.1016/S0022-1031(02)00513-9. [34](#), [129](#)
- [373] R. B. McIntyre, C. G. Lord, D. M. Gresky, L. L. T. Eyck, and C. F. Bond. A Social Impact Trend in the Effects of Role Models on Alleviating Women’s Mathematics Stereotype Threat. *Current Research in Social Psychology*, 10(9):1–26, 2005. [34](#), [184](#), [188](#)
- [374] R. Mehta and R. Zhu. Blue or Red? Exploring the Effect of Color on Cognitive Task Performances. *Science*, 323(February):1226–1229, 2008. ISSN 00989258. doi: 10.1126/science.1169144. [25](#), [93](#), [147](#), [148](#), [154](#), [209](#)
- [375] B. P. Meier, P. R. D’Agostino, A. J. Elliot, M. a. Maier, and B. M. Wilkowski. Color in context: Psychological context moderates the influence of red on approach- and avoidance-motivated behavior. *PLoS ONE*, 7(7):1–5, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0040333. [154](#)
- [376] M. Meier, R. A. Hill, A. J. Elliot, and R. Barton. Color in Achievement Contexts in Humans. *Handbook of Color Psychology*, 44(February):0–103, 2015. ISSN 1881-8323. doi: 10.1063/1.2756072. URL <http://dx.doi.org/10.1016/j.worlddev.2005.07.015>. [25](#), [93](#), [147](#), [148](#), [153](#), [154](#), [193](#), [209](#)

- [377] C. Midgley. *Goals, goal structures, and patterns of adaptive learning*. 2002. ISBN 0-8058-3884-8. URL <http://search.ebscohost.com/login.aspx?direct=true{%&}db=psyh{%&}AN=2002-12839-000{%&}lang=ja{%&}site=ehost-live>. 44
- [378] J. P. Mitchell, C. N. Macrae, and M. R. Banaji. Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others. *Neuron*, 50(4):655–663, 2006. ISSN 08966273. doi: 10.1016/j.neuron.2006.03.040. 31, 132, 189
- [379] H. Mitterer, J. M. Horschig, J. Müsseler, and A. Majid. The influence of memory on perception: it's not what things look like, it's what you call them. *Journal of experimental psychology. Learning, memory, and cognition*, 35(6):1557–1562, 2009. ISSN 0278-7393. doi: 10.1037/a0017019. 147
- [380] D. Mobbs, R. Yu, M. Meyer, L. Passamonti, B. Seymour, A. J. Calder, S. Schweizer, C. D. Frith, and T. Dalgleish. A key role for similarity in vicarious reward. *Science*, 324(5929):900, 2009. ISSN 1095-9203. doi: 10.1126/science.1170539. URL <http://www.sciencemag.org.ezproxy.is.ed.ac.uk/content/324/5929/900.full>. 31, 132, 136, 189
- [381] R. J. Moffatt, L. F. Chitwood, and K. D. Biggerstaff. The influence of verbal encouragement during assessment of maximal oxygen uptake. *Journal of Sports Medicine and Physical Fitness*, 34(1):45–49, 1994. ISSN 0022-4707. 155, 164
- [382] A. C. Moller, A. J. Elliot, and M. a. Maier. Basic hue-meaning associations. *Emotion (Washington, D.C.)*, 9(6):898–902, 2009. ISSN 1528-3542. doi: 10.1037/a0017811. 148
- [383] J. D. Mollon. "Tho' she kneel'd in that place where they grew..." The uses and origins of primate colour vision. *The Journal of experimental biology*, 146:21–38, 1989. ISSN 0022-0949. 147
- [384] E. K. Molloy and D. Boud. Handbook of Research on Educational Communications

- and Technology. pages 413–424, 2014. doi: 10.1007/978-1-4614-3185-5. URL <http://link.springer.com/10.1007/978-1-4614-3185-5>. 26, 94, 155
- [385] M. Montola, T. Nummenmaa, A. Lucero, M. Boberg, and H. Korhonen. Applying game achievement systems to enhance user experience in a photo sharing service. *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*, 4(46):94–97, 2009. doi: 10.1145/1621841.1621859. URL <http://portal.acm.org/citation.cfm?id=1621841.1621859{%&}coll=GUIDE{%&}dl=GUIDE{%&}CFID=103096974{%&}CFTOKEN=12484716>. 187
- [386] Y. Moon. Intimate Exchanges: Using Computers to Elicit Self-Disclosure From Consumers. *Journal of Consumer Research*, 26(4):323–339, 2000. ISSN 0093-5301. doi: 10.1086/209566. URL <http://www.jstor.org/stable/10.1086/209566>. 31, 131, 132, 136, 189
- [387] M. M. Moretti and E. T. Higgins. Relating self-discrepancy to self-esteem: The contribution of discrepancy beyond actual-self ratings. *Journal of Experimental Social Psychology*, 26(2):108–123, 1990. 207
- [388] M. Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970. ISSN 10709932. doi: 10.1109/MRA.2012.2192811. 24, 92
- [389] J. F. Morie and G. Verhulsdonck. Body/Persona/Action! Emerging Non-anthropomorphic Communication and Interaction in Virtual Worlds. *Advances on Computer Entertainment Technology*, page 365, 2008. doi: 10.1145/1501750.1501837. 95
- [390] B. B. Morrison and B. DiSalvo. Khan academy gamifies computer science. In *Proceedings of the 45th ACM technical symposium on Computer science education*, pages 39–44. ACM, 2014. 184, 187, 207
- [391] E. Moyer-Gusé, A. H. Chung, and P. Jain. Identification with characters and discussion

- of taboo topics after exposure to an entertainment narrative about sexual health. *Journal of Communication*, 61(3):387–406, 2011. 189
- [392] Mozilla Foundation, Peer 2 Peer University, and MacArthur Foundation. Open Badges for Lifelong Learning. 2013. 186
- [393] C. M. Mueller and C. S. Dweck. Praise for intelligence can undermine children’s motivation and performance. *Journal of personality and social psychology*, 75(1): 33–52, 1998. ISSN 0022-3514. doi: 10.1037/0022-3514.75.1.33. 156
- [394] J. Mumm and B. Mutlu. Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. *Computers in Human Behavior*, 27(5):1643–1650, 2011. ISSN 07475632. doi: 10.1016/j.chb.2011.02.002. URL <http://linkinghub.elsevier.com/retrieve/pii/S0747563211000355>. 164
- [395] L. Nacke, A. Drachen, K. Kuikkaniemi, J. Niesenhaus, H. J. Korhonen, W. M. Hoogen, K. Poels, W. A. IJsselsteijn, and Y. A. W. De Kort. Playability and player experience research. In *Proceedings of DiGRA 2009: Breaking New Ground: Innovation in Games, Play, Practice and Theory*. DiGRA, 2009. 78
- [396] A. Namousi and D. S. Kohl. Crowdsourcing in Video Games: The Motivational Factors of the Crowd. *The Computer Games Journal*, pages 1–15, 2016. 39
- [397] S. Narciss. Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. *Digital Education Review*, 23(1):7–26, 2013. ISSN 20139144. 164
- [398] N. S. Nasir. Identity, Goals, and Learning: Mathematics in Cultural Practice, 2002. ISSN 1098-6065.
- [399] R. Ng and R. Lindgren. Examining the effects of avatar customization and narrative on engagement and learning in video games. *Proceedings of CGAMES 2013 USA - 18th International Conference on Computer Games: AI, Animation, Mobile, Interactive*

- Multimedia, Educational and Serious Games*, pages 87–90, 2013. doi: 10.1109/CGames.2013.6632611. 189
- [400] S. Niedenthal. Dynamic Lighting for Tension in Games. *Game Studies*, 2007. 148
- [401] A. I. Nordin, A. Denisova, and P. Cairns. Too Many Questionnaires: Measuring Player Experience Whilst Playing Digital Games. *Seventh York Doctoral Symposium on Computer Science & Electronics 69*, 2014. 164
- [402] K. L. Norman. GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers. *Interacting with Computers*, 25(4):278–283, mar 2013. ISSN 0953-5438. doi: 10.1093/iwc/iwt009. URL <http://iwc.oxfordjournals.org/content/25/4/278.abstract>. 79, 164
- [403] J. J. Odell, H. V. D. Parunak, and M. Fleischer. The Role of Roles in Designing Effective Agent Organizations. *Software Engineering for Large-Scale Multi-Agent Systems*, pages 27–38, 2003. ISSN 03029743. doi: 10.1177/0013916507311547. 137
- [404] C. O’Donovan, E. Hirsch, E. Holohan, I. McBride, R. McManus, and J. Hussey. Energy expended playing Xbox Kinect And Wii Games: A preliminary study comparing single and multiplayer modes. *Physiotherapy (United Kingdom)*, 98(3): 224–229, 2012. ISSN 00319406. doi: 10.1016/j.physio.2012.05.010. 141
- [405] J. Orkin and D. Roy. The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development*, 3(December): 39–60, 2007. URL <http://media.mit.edu/cogmac/publications/Orkin{ }JoGD07{ }inpress.pdf>. 154
- [406] E. O’Rourke and K. Haimovitz. Brain points: a growth mindset incentive structure boosts persistence in an educational game. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI ’14*, pages 3339–3348, 2014. URL <http://dl.acm.org/citation.cfm?id=2557157>. 155

- [407] M. Overmars. Teaching computer science through game design. *Computer*, 37(4): 81–83, 2004. 45
- [408] D. Ozdemir and P. Doolittle. Revisiting the Seductive Details Effect in Multimedia Learning: Context-Dependency of Seductive Details. *Journal of Educational Multimedia and Hypermedia*, 24(2):101–119, 2015. ISSN ISSN-1055-8896. 173
- [409] S. P. Walz and S. Deterding. *The Gameful World - Approaches, Issues, Applications*. 2015. ISBN 9788578110796. doi: 10.1017/CBO9781107415324.004. URL <http://www.tandfonline.com/doi/full/10.1080/00140139.2015.1067048>. 184, 187
- [410] F. Pajares. Self-Efficacy Beliefs in Academic Settings. *Review of Educational Research*, 66(4):543–578, 1996. ISSN 00346543, 19351046. doi: 10.3102/00346543066004543. URL <http://www.jstor.org.proxy2.library.illinois.edu/stable/1170653>. 77, 172
- [411] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. 2010. 80
- [412] S. Papert. *Mindstorms*, volume 15. 1980. doi: 10.1111/j.1467-8721.2006.00431.x. URL <http://cdp.sagepub.com/content/15/4/177.abstract>{%}5Cn<http://cdp.sagepub.com/content/15/4/177.full.pdf>{%}5Cn<http://cdp.sagepub.com/content/15/4/177.short>. 40
- [413] S. Papert. *Mindstorms*, 1993. 37
- [414] S. Papert. An exploration in the space of mathematics educations. *International Journal of Computers for Mathematical Learning*, 1(1), 1996. ISSN 1382-3892. doi: 10.1007/BF00191473. URL <http://link.springer.com/10.1007/BF00191473>. 40

- [415] S. Papert. Papert on piaget. *Time magazine's special issue on "The Century's Greatest Minds"*, page 105, 1999. 37
- [416] S. Papert and I. Harel. Situating Constructionism. *Constructionism*, 1991. 36, 184
- [417] B. Park, R. Moreno, T. Seufert, and R. Brünken. Does cognitive load moderate the seductive details effect? A multimedia study. *Computers in Human Behavior*, 27(1): 5–10, 2011. ISSN 07475632. doi: 10.1016/j.chb.2010.05.006. 167, 172, 173
- [418] B. Park, T. Flowerday, and R. Brünken. Cognitive and affective effects of seductive details in multimedia learning. *Computers in Human Behavior*, 44:267–278, 2015. ISSN 07475632. doi: 10.1016/j.chb.2014.10.061. URL <http://dx.doi.org/10.1016/j.chb.2014.10.061>. 172
- [419] B.-W. Park and K. C. Lee. Exploring the value of purchasing online game items. *Computers in Human Behavior*, 27(6):2178–2185, 2011. 189
- [420] L. E. Parker and M. R. Lepper. Effects of Fantasy Contexts on Children's Learning and Motivation. *Journal of Personality and Social Psychology*, 62(4):625–633, 1992. ISSN 0022-3514. doi: 10.1037/0022-3514.62.4.625. URL <http://sfx.bibl.ulaval.ca:9003/sfx{ }local?ctx{ }ver=Z39.88-2004{&}ctx{ }enc=info:ofi/enc:UTF-8{&}ctx{ }tim=2013-05-26T08{ }3A39{ }3A07IST{&}url{ }ver=Z39.88-2004{&}url{ }ctx{ }fmt=infofi/fmt:kev:mtx:ctx{&}rfr{ }id=info:sid/primo.exlibrisgroup.com:primo3-Article-apa{ }articles{&}rft{ }val{ }f>. 173
- [421] G. Parmentier and R. Gandia. Managing sustainable innovation with a user community toolkit: the case of the video game Trackmania. *Creativity and Innovation Management*, 22(2):195–208, 2013. 39
- [422] E. a. Patall, H. Cooper, and J. C. Robinson. The effects of choice on intrinsic motivation and related outcomes: a meta-analysis of research findings. *Psychological*

- Bulletin*, 134(2):270–300, 2008. ISSN 0033-2909. doi: 10.1037/0033-2909.134.2.270. [168](#)
- [423] H. Patel, J. Andrade, and M. Blades. Children’s incidental learning of the colors of objects and clothing, 2001. ISSN 08852014. [147](#)
- [424] V. Payen, A. J. Elliot, S. a. Coombes, A. Chalabaev, J. Brisswalter, and F. Cury. Viewing red prior to a strength test inhibits motor output. *Neuroscience Letters*, 495: 44–48, 2011. ISSN 03043940. doi: 10.1016/j.neulet.2011.03.032. [147](#), [148](#)
- [425] J. Pena, J. T. Hancock, and N. a. Merola. The Priming Effects of Avatars in Virtual Settings. *Communication Research*, 36(6):838–856, 2009. ISSN 0093-6502. doi: 10.1177/0093650209346802. [148](#)
- [426] W. Peng and J. Crouse. Playing in parallel: the effects of multiplayer modes in active video game on motivation and physical exertion. *Cyberpsychology, behavior and social networking*, 16(6):423–7, 2013. ISSN 2152-2723. doi: 10.1089/cyber.2012.0384. URL <http://www.ncbi.nlm.nih.gov/pubmed/23509986>. [141](#)
- [427] J. Pennebaker and C. Chung. The Development and Psychometric Properties of LIWC2007. *Austin, TX, LIWC. . . .*, pages 1–22, 2007. [101](#)
- [428] J. W. Pennebaker and L. A. King. Linguistic styles: language use as an individual difference. *Journal of pers. and social psych.*, 1999. ISSN 0022-3514. doi: 10.1037/0022-3514.77.6.1296. [101](#)
- [429] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology*, 54: 547–577, 2003. ISSN 0066-4308. doi: 10.1146/annurev.psych.54.101601.145041. [101](#)
- [430] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The Development and Psychometric Properties of LIWC2015. 2015. [152](#)

- [431] J. Piaget. *Piaget and His School*. Springer Berlin Heidelberg, 1976. ISBN 978-3-540-07248-5. [36](#)
- [432] M. Piccioni, C. Estler, and B. Meyer. SPOC-supported introduction to programming. In *Proceedings of the 2014 conference on Innovation & technology in computer science education*, pages 3–8. ACM, 2014. [184](#), [187](#)
- [433] J. L. Plass, S. Heidig, E. O. Hayward, B. D. Homer, and E. Um. Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction*, 29:128–140, 2014. ISSN 09594752. doi: 10.1016/j.learninstruc.2013.02.006.
- [434] J. L. Plass, P. A. O’Keefe, M. L. Biles, J. Frye, and B. D. Homer. Motivational and Cognitive Impact of Badges in Games for Learning. (October 2015), 2014. doi: 10.13140/2.1.3209.9842. [187](#)
- [435] K. Poels, Y. De Kort, and W. Ijsselsteijn. It is always a lot of fun!: exploring dimensions of digital game experience using focus group methodology. In *Proceedings of the 2007 conference on Future Play*, pages 83–89. ACM, 2007. [78](#)
- [436] H. Postigo. From Pong to Planet Quake: Post-Industrial Transitions from Leisure to Work. *Information, Communication & Society*, 6(4):593–607, 2003. ISSN 1369-118X. [39](#)
- [437] H. Postigo. Modification. In *Debugging Game History: A Critical Lexicon*, page 325. MIT Press, 2016. [38](#)
- [438] J. A. Pratt, K. Hauser, Z. Ugray, and O. Patterson. Looking at human-computer interface design: Effects of ethnicity in computer agents. *Interacting with Computers*, 19(4):512–523, 2007. ISSN 09535438. [31](#), [130](#), [132](#), [136](#), [138](#)
- [439] K. Pravossoudovitch, F. Cury, S. G. Young, and A. J. Elliot. Is red the colour of danger? Testing an implicit red-danger association. *Ergonomics*, 57(4):503–

- 10, 2014. ISSN 1366-5847. doi: 10.1080/00140139.2014.889220. URL <http://www.ncbi.nlm.nih.gov/pubmed/24588355>. 148
- [440] J. S. Prook, D. P. Janssen, and S. Gualeni. The Negative Effects of Praise and Flattery. *FDG*, 2015. 164
- [441] A. K. Przybylski, N. Weinstein, K. Murayama, M. F. Lynch, and R. M. Ryan. The ideal self at play: The appeal of video games that let you be all you can be. *Psychological science*, 23(1):69–76, 2012. 190
- [442] H. S. Raffle, A. J. Parkes, and H. Ishii. Topobo: a constructive assembly system with kinetic memory. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 647–654. ACM, 2004. 38
- [443] V. Ramalingam and S. Wiedenbeck. Development and Validation of Scores on a Computer Programming Self-Efficacy Scale and Group Analyses of Novice Programmer Self-Efficacy. *Journal of Educational Computing Research*, 19(4):367–381, 1998. ISSN 0735-6331. doi: 10.2190/C670-Y3C8-LTJ1-CT3P. 169, 197
- [444] A. Ramaprasad. On the definition of feedback. *Behavioral Science*, 28(1):4–13, 1983. ISSN 1099-1743. doi: 10.1002/bs.3830280103. URL <http://onlinelibrary.wiley.com/doi/10.1002/bs.3830280103/abstract>. 26, 94, 155
- [445] R. Ratan and Y. J. Sah. Leveling up on stereotype threat: The role of avatar customization and avatar embodiment. *Computers in Human Behavior*, 50:367–374, 2015. ISSN 07475632. doi: 10.1016/j.chb.2015.04.010. URL <http://linkinghub.elsevier.com/retrieve/pii/S0747563215002940>. 32, 189, 214
- [446] R. Ratan and H. Y. Tsai. Dude, where’s my avacar? A mixed-method examination of communication in the driving context. *Pervasive and Mobile Computing*, 14: 112–128, 2014. ISSN 15741192. doi: 10.1016/j.pmcj.2014.05.011. URL <http://dx.doi.org/10.1016/j.pmcj.2014.05.011>.
- [447] J. Reeve, B. C. Olson, and S. G. Cole. Motivation and performance: Two con-

- sequences of winning and losing in competition. *Motivation and Emotion*, 9(3): 291–298, 1985. ISSN 01467239. doi: 10.1007/BF00991833. 165
- [448] J. Reeve, B. C. Olson, and S. G. Cole. Intrinsic motivation in competition: The intervening role of four individual differences following objective competence information. *Journal of Research in Personality*, 21(2):148–170, 1987. ISSN 0092-6566. doi: [http://dx.doi.org/10.1016/0092-6566\(87\)90004-3](http://dx.doi.org/10.1016/0092-6566(87)90004-3). URL <http://www.sciencedirect.com/science/article/pii/S0092656687900043>. 165
- [449] A. Repenning. Agentsheets: a tool for building domain-oriented visual programming environments. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 142–143. ACM, 1993. 45
- [450] L. B. Resnick. Learning in school and out. *Educational Researcher*, 16(9):13–20, 1987. ISSN 0013189X. doi: 10.3102/0013189X029002004. URL <http://www.jstor.org/stable/1175725>. 167
- [451] M. Resnick. Still a badge skeptic. *Retrieved June, 2:2015*, 2012. 186
- [452] M. Resnick and J. Maloney. Scratch: programming for all. *Communications of the ...*, 2009. URL <http://dl.acm.org/citation.cfm?id=1592779>. 3, 38, 45, 184
- [453] M. Resnick, M. Flanagan, C. Kelleher, M. MacLaurin, Y. Ohshima, K. Perlin, and R. Torres. Growing Up Programming Democratizing the Creation of Dynamic, Interactive Media. *CHI EA '09 Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, April 4-9, 2009, Boston, Massachusetts, USA*, pages 3293–3296, 2009. doi: 10.1145/1520340.1520472. 167, 172
- [454] M. Restivo and A. Van De Rijt. Experimental study of informal rewards in peer production. *PloS one*, 7(3):e34358, 2012. 187

- [455] G. D. Rey. A review of research and a meta-analysis of the seductive detail effect, 2012. ISSN 1747938X. 26, 94, 166, 167
- [456] M. Rice, R. Koh, Q. Lui, Q. He, M. Wan, V. Yeo, J. Ng, and W. P. Tan. Comparing avatar game representation preferences across three age groups. *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, page 1161, 2013. doi: 10.1145/2468356.2468564. URL <http://dl.acm.org/citation.cfm?doid=2468356.2468564>. 201
- [457] L. P. Rieber. Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research and Development*, 44(2):43–58, 1996. ISSN 1042-1629. doi: 10.1007/BF02300540. 167, 173
- [458] S. Rigby and R. M. Ryan. *Glued to games: How video games draw us in and hold us spellbound: How video games draw us in and hold us spellbound*. ABC-CLIO, 2011. 78, 187, 197
- [459] M. Robertson. Can't play, won't play. *Hide & Seek*, 6:2010, 2010. 186
- [460] Roblox. We're Building the Future of Entertainment, 2016. URL <http://corp.roblox.com/careers>. 38
- [461] I. Roll, V. Aleven, B. M. McLaren, and K. R. Koedinger. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2):267–280, 2011. ISSN 09594752. doi: 10.1016/j.learninstruc.2010.07.004. URL <http://linkinghub.elsevier.com/retrieve/pii/S0959475210000538>. 164
- [462] D. H. Rose and A. Meyer. Teaching Every Student in the Digital Age: Universal Design for Learning. *Ericedgov*, page 216, 2002. doi: 10.1007/s11423-007-9056-3. URL <http://www.cast.org/teachingeverystudent/ideas/tes/>. 171, 207

- [463] K. J. Rose, M. Koenig, and F. Wiesbauer. Evaluating success for behavioral change in diabetes via mHealth and gamification: MySugr's keys to retention and patient engagement. *Diabetes Technology & Therapeutics*, 15:A114, 2013. 186
- [464] R. B. Rosenberg-Kima, E. A. Plant, C. E. Doerr, and A. Baylor. The influence of computer-based model's race and gender on female students' attitudes and beliefs towards engineering. *Journal of Engineering Education*, pages 35–44, 2010. ISSN 10694730. doi: 10.1002/j.2168-9830.2010.tb01040.x. 31, 130, 132, 136, 138
- [465] E. Rowland, C. H. Skinner, K. Davis-Richards, R. Saudargas, and D. H. Robinson. An Investigation of Placement and Type of Seductive Details: The Primacy Effect of Seductive Details on Text Recall. 15(2):80–90, 2008. 172
- [466] Royal Society. Shut Down or Restart? The Way Forward for Computing in UK Schools, 2012. URL <https://advancesinap.collegeboard.org/stem/computer-science-principles>. 40
- [467] R. Rughinis. Talkative objects in need of interpretation: Re-thinking digital badges in education. *CHI '13 extended abstracts on human factors in computing systems*, pages 2099–2108, 2013. doi: 10.1145/2468356.2468729. URL <http://altchi.org/2013/submissions/submission{ }razvan.rughinis{ }0.pdf>.
- [468] J. A. Ruipérez-Valiente, P. J. Muñoz-Merino, and C. D. Kloos. An analysis of the use of badges in an educational experiment. *Proceedings - Frontiers in Education Conference, FIE*, 2016-Novem:1–8, 2016. ISSN 15394565. doi: 10.1109/FIE.2016.7757424. 184, 187
- [469] R. Rummer, J. Schweppe, A. Fürstenberg, T. Seufert, and R. Brünken. Working memory interference during processing texts and pictures: Implications for the explanation of the modality effect. *Applied Cognitive Psychology*, 24(2):164–176, 2010. ISSN 08884080. doi: 10.1002/acp.1546. 167
- [470] J. B. Russ, R. C. Gur, and W. B. Bilker. Validation of affective and neutral sentence

- content for prosodic testing. *Behavior research methods*, 40(4):935–9, 2008. ISSN 1554-351X. doi: 10.3758/BRM.40.4.935. URL <http://www.ncbi.nlm.nih.gov/pubmed/19001384>. 156
- [471] A. M. Rutchick, M. L. Slepian, and B. D. Ferris. The pen is mightier than the word: Object priming of evaluative standards. *European Journal of Social Psychology*, 40: 704–708, 2010. ISSN 00462772. doi: 10.1002/ejsp.753. 148
- [472] R. M. Ryan and E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American psychologist*, 55(1): 68–78, 2000. ISSN 0003-066X. doi: 10.1037/0003-066X.55.1.68. 168, 197
- [473] R. M. Ryan and E. L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67, 2000. 186
- [474] R. M. Ryan, C. S. Rigby, and A. Przybylski. The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion*, 30 (4):344–360, 2006. ISSN 0146-7239. doi: 10.1007/s11031-006-9051-8. URL <http://link.springer.com/10.1007/s11031-006-9051-8>. 21, 78, 169, 197, 201
- [475] M. Rymaszewski. *Second life: The official guide*, volume 2. John Wiley & Sons, 2007. 39
- [476] P. R. Sackett, C. M. Hardison, and M. J. Cullen. On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. *American Psychologist*, 59(1):7, 2004. 35
- [477] P. R. Sackett, C. M. Hardison, and M. J. Cullen. On the Value of Correcting Mischaracterizations of Stereotype Threat Research. 2004. 35
- [478] D. R. Sadler. Formative Assessment and the design of instructional systems. *Instructional Science*, 18:119–144, 1989. ISSN 0969-594X. doi: 10.1007/BF00117714. 26, 94, 155

- [479] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, 69:371–380, 2017. ISSN 07475632. doi: 10.1016/j.chb.2016.12.033. URL <http://dx.doi.org/10.1016/j.chb.2016.12.033>. 184
- [480] K. Salen. *Quest to learn: Developing the school for digital kids*. MIT Press, 2011. 186
- [481] C. a. Sanchez and J. Wiley. An examination of the seductive details effect in terms of working memory capacity. *Memory & cognition*, 34(2):344–55, 2006. ISSN 0090-502X. doi: 10.3758/BF03193412. URL <http://www.ncbi.nlm.nih.gov/pubmed/16752598>. 167, 172
- [482] J. Sanders. By the Numbers: 10 Stats on the Growth of Gamification, 2015. URL <http://www.gamesandlearning.org/2015/04/27/by-the-numbers-10-stats-on-the-growth-of-gamification/>. 19
- [483] K. Scheiter, P. Gerjets, T. Huk, B. Imhof, and Y. Kammerer. The effects of realism in learning with dynamic visualizations. *Learning and Instruction*, 19(6):481–494, 2009. ISSN 09594752. doi: 10.1016/j.learninstruc.2008.08.001. 173
- [484] T. Schmader and M. Johns. Converging evidence that stereotype threat reduces working memory capacity. *Journal of personality and social psychology*, 85(3):440, 2003. 34
- [485] T. Schmader, M. Johns, and C. Forbes. An integrated process model of stereotype threat effects on performance. *Psychological Review*, 115(2):336–356, 2008. ISSN 1939-1471. doi: 10.1037/0033-295X.115.2.336. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.115.2.336>. 34
- [486] N. Schüll and N. Schull. Addiction by design : machine gambling in Las Vegas.

- The Annals of the American Academy of Political and Social Science*, 597(1):65 – 81, 2012. ISSN 0002-7162. doi: 10.2307/j.ctt12f4d0. 165
- [487] E. Schweikardt and M. D. Gross. roBlocks: a robotic construction kit for mathematics and science education. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 72–75. ACM, 2006. 38
- [488] M. J. Scott and G. Ghinea. Integrating Fantasy Role-Play Into the Programming Lab. *Proceeding of the 44th ACM Technical Symposium on Computer Science Education - SIGCSE '13*, page 119, 2013. doi: 10.1145/2445196.2445237. URL <http://dl.acm.org/citation.cfm?id=2445196.2445237>. 26, 94, 154, 166, 167, 172
- [489] K. Seaborn and D. I. Fels. Gamification in theory and action: A survey. *International Journal of Human Computer Studies*, 74:14–31, 2015. ISSN 10959300. doi: 10.1016/j.ijhcs.2014.09.006. 184, 186, 187
- [490] A. Sfard. On Two Metaphors for learning and the Dangers of Choosing Just One. *Educational Researcher*, 27(2):4–13, 1998. ISSN 0013-189X. doi: 10.3102/0013189X027002004. 167
- [491] D. W. Shaffer. Epistemic frames for epistemic games. *Computers & Education*, 46(3):223–234, 2006. ISSN 03601315. doi: 10.1016/j.compedu.2005.11.003. 37, 167
- [492] B. A. Sheil. Teaching procedural literacy (Presentation Abstract). In *Proceedings of the ACM 1980 annual conference*, pages 125–126. ACM, 1980. 40
- [493] J. Shi, C. Zhang, and F. Jiang. Does red undermine individuals' intellectual performance? A test in China. *International Journal of Psychology*, 50(1):81–84, 2015. ISSN 00207594. doi: 10.1002/ijop.12076. URL <http://doi.wiley.com/10.1002/ijop.12076>. 25, 93, 147, 148, 154
- [494] L. Shi, A. I. Cristea, J. G. K. Foss, D. A. Qudah, and A. Qaffas. A social personalized

- adaptive e-learning environment: a case study in topolor. *IADIS International Journal on WWW/Internet*, 11(3):13–34, 2013. 44
- [495] M. Shih, T. L. Pittinsky, and N. Ambady. Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance, 1999. ISSN 0956-7976. 33
- [496] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420–428, 1979. ISSN 0033-2909. doi: 10.1037/0033-2909.86.2.420. 156, 193
- [497] V. J. Shute. Focus on Formative Feedback. *Review of Educational Research*, 78(1): 153–189, 2008. ISSN 0034-6543. doi: 10.3102/0034654307313795. URL <http://rer.sagepub.com/cgi/doi/10.3102/0034654307313795>. 156, 164
- [498] A. Sipitakiat, P. Blikstein, and D. P. Cavallo. GoGo board: augmenting programmable bricks for economically challenged audiences. In *Proceedings of the 6th international conference on Learning sciences*, pages 481–488. International Society of the Learning Sciences, 2004. 45
- [499] D. Smahel, L. Blinka, and O. Ledabyl. Playing MMORPGs: connections between addiction and identifying with a character. *Cyberpsychology & Behavior*, 11(6): 715–718, 2008. ISSN 1094-9313. doi: 10.1089/cpb.2007.0210. 184, 189
- [500] A. R. B. Soutter and M. Hitchens. The relationship between character identification and flow state within video games. *Computers in Human Behavior*, 55(December 2015):1030–1038, 2016. ISSN 07475632. doi: 10.1016/j.chb.2015.11.012. URL <http://dx.doi.org/10.1016/j.chb.2015.11.012>. 184, 188, 189
- [501] SRI International. ECS Assessments. URL <https://csforallteachers.org/blog/release-of-the-ecs-assessments-cumulative-units-1-4-assessments-available-now>.
- [502] M. Stange, C. Graydon, and M. J. Dixon. “It was that close”: Investigating Players’ Reactions to Losses, Wins, and Near-Misses on Scratch Cards. *Journal of*

- Gambling Studies*, 2015. ISSN 1573-3602. doi: 10.1007/s10899-015-9538-x. URL <http://link.springer.com/10.1007/s10899-015-9538-x>. 165
- [503] C. Steele and J. Aronson. Stereotype Threat and the Intellectual Test Performance of African Americans. *Journal of personality and social psychology*, 1995. 19, 24, 32, 44, 92
- [504] C. M. Steele. Whistling Vivaldi and other clues to how stereotypes affect us. In *Whistling Vivaldi*. 2010. ISBN 9780393062496 (hbk.). doi: 10.1126/science.1194619. 33, 35
- [505] S. T. Steinemann, E. D. Mekler, and K. Opwis. Increasing donating behavior through a game for change: The role of interactivity and appreciation. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, pages 319–329. ACM, 2015.
- [506] J. Steinke, B. Applegate, M. Lapinski, L. Ryan, and M. Long. Gender Differences in Adolescents' Wishful Identification With Scientist Characters on Television. *Science Communication*, 34(2):163–199, 2012. ISSN 1075-5470. doi: 10.1177/1075547011410250. URL <http://scx.sagepub.com>. 189
- [507] J. N. Stinson, L. A. Jibb, C. Nguyen, P. C. Nathan, A. M. Maloney, L. L. Dupuis, J. T. Gerstle, B. Alman, S. Hopyan, C. Strahlendorf, and Others. Development and testing of a multidimensional iPhone pain assessment application for adolescents with cancer. *Journal of medical Internet research*, 15(3), 2013. 186
- [508] G. Stoet and D. C. Geary. Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, 16(1):93–102, 2012. ISSN 1939-1552. doi: 10.1037/a0026617. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0026617>. 35
- [509] K. T. Stolee and T. Fristoe. Expressing computer science concepts through Kodu

- game lab. In *Proceedings of the 42nd ACM technical symposium on Computer science education*, pages 99–104. ACM, 2011. 45
- [510] J. Stone, W. Perry, and J. M. Darley. "White Men Can't Jump": Evidence for the Perceptual Confirmation of Racial Stereotypes Following a Basketball Game. *Basic and Applied Social Psychology*, 19(3):291–306, 1997. 33
- [511] H. Tajfel. Experiments in intergroup discrimination. *Scientific American*, 223(5):96–102, 1970. ISSN 0036-8733. doi: 10.1038/scientificamerican1170-96. 33
- [512] C.-I. Teng. Impact of avatar identification on online gamer loyalty: Perspectives of social identity and social capital theories. *International Journal of Information Management*, 37(6):601–610, 2017. ISSN 02684012. doi: 10.1016/j.ijinfomgt.2017.06.006. URL <http://linkinghub.elsevier.com/retrieve/pii/S0268401216308507>. 184, 189
- [513] W. Thalheimer. Bells, whistles, neon, and purple prose: When interesting words, sounds, and visuals hurt learning and performance – a review of the seductive-augmentation research. pages 1–29, 2004. 26, 94, 166, 167
- [514] The TurkPrime Team. The New New Demographics on Mechanical Turk: Is there Still a Gender Gap? URL <http://blog.turkprime.com/2015/03/the-new-new-demographics-on-mechanical.html>. 80
- [515] J. Thom, D. Millen, and J. DiMicco. Removing gamification from an enterprise SNS. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, pages 1067–1070. ACM, 2012. 186
- [516] M. Thomsen. "Super Mario Maker" is an engine for circulating horrible new "Mario" levels, 2015. URL https://www.washingtonpost.com/news/comic-riffs/wp/2015/09/15/super-mario-maker-is-an-engine-for-circulating-horrible-new-mario-levels/?utm_{_}term=.fdcl1e806bdc3. 209

- [517] W. G. Tierney, Z. B. Corwin, T. Fullerton, and G. Ragusa. *Postsecondary Play: The Role of Games and Social Media in Higher Education*. JHU Press, 2014. ISBN 142141306X. URL <https://books.google.com/books?hl=en&lr=&id=O16RAwAAQBAJ&pgis=1>. 26, 94, 166
- [518] S. Trepte and L. Reinecke. Avatar creation and video game enjoyment. *Journal of Media Psychology*, 2010. 189, 190
- [519] T. Trust, R. W. Maloy, and S. Edwards. Learning through Making: Emerging and Expanding Designs for College Classes. *Tech Trends*, 2017. ISSN 87563894. doi: 10.1007/s11528-017-0214-0. 184, 187
- [520] S. Turkay. The Effects of Avatar-based Customization on Player Identification in Extended MMO Gameplay. *Games, Learning and Society Conference*, 6(March): 227–234, 2014. ISSN 1942-3888. doi: 10.4018/ijgcms.2014010101. 32, 132, 190
- [521] S. Turkay and C. K. Kinzer. *The Relationship between Avatar-Based Customization, Player Identification, and Motivation*. ISBN 9781522518174. doi: 10.4018/978-1-5225-1817-4.ch003. URL <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-5225-1817-4.ch003>. 184, 189, 199, 214
- [522] T. Tyler-Wood, A. Ellison, O. Lim, and S. Periathiruvadi. Bringing up girls in science (BUGS): The effectiveness of an afterschool environmental science program for increasing female students' interest in science careers. *Journal of Science Education and Technology*, 21(1):46–55, 2012.
- [523] S. E. Ural and S. Yilmazer. The architectural color design process: An evaluation of sequential media via semantic ratings. *Color Research and Application*, 2010. 148
- [524] Valve. Dota 2 Workshop - Item FAQ, 2016. URL <https://support.steampowered.com/kb/5538-WPZC-9529/dota-2-workshop-item-faq{#}publishingprocess>. 39

- [525] Valve. Steam Workshop, 2017. URL <http://steamcommunity.com/workshop/>. 39
- [526] J. Van Looy, C. Courtois, M. De Vocht, and L. De Marez. Player Identification in Online Games: Validation of a Scale for Measuring Identification in MMOGs. *Media Psychology*, 15(2):197–221, 2012. ISSN 1521-3269. doi: 10.1080/15213269.2012.674917. URL <http://online.liebertpub.com/doi/abs/10.1089/g4h.2014.0010{%}5Cnhttp://www.tandfonline.com/doi/abs/10.1080/15213269.2012.674917>. 80, 189, 190, 198
- [527] E. A. Van Reijmersdal, J. Jansz, O. Peters, and G. Van Noort. Why girls go pink: Game character identification and game-players’ motivations. *Computers in Human Behavior*, 29(6):2640–2649, 2013. 189
- [528] A. Vasalou, A. N. Joinson, and J. Pitt. Constructing my online self: avatars that increase self-focused attention. *Proceedings of ACM CHI 2007 Conference on Human Factors in Computing Systems*, 1:445–448, 2007. doi: 10.1145/1240624.1240696. URL <http://doi.acm.org/10.1145/1240624.1240696>. 31, 131, 132, 136, 189
- [529] S. Veeragoudar Harrell. Representation, medium, and agency in mathematics practice: Toward the development of a model of mathematical agency. *American Education Research Association*, 2007. 19
- [530] M. Virvou, G. Katsionis, and K. Manos. Combining software games with education: Evaluation of its educational effectiveness, 2005. ISSN 11763647. 173
- [531] T. F. Waddell, S. S. Sundar, and J. Auriemma. Can customizing an avatar motivate exercise intentions and health behaviors among those with low health ideals? *Cyberpsychology, Behavior, and Social Networking*, 18(11):687–690, 2015. 189
- [532] M. Wadhwa and J. C. Kim. Can a Near Win Kindle Motivation? The Impact of

- Nearly Winning on Motivation for Unrelated Rewards. *Psychological Science*, 26(6): 701–708, 2015. ISSN 0956-7976. doi: 10.1177/0956797614568681. URL <http://pss.sagepub.com/lookup/doi/10.1177/0956797614568681>. 165
- [533] Walker. Developer at Valve [Interview], 2013. 39
- [534] V. Walkerdine. Redefining the subject in situated cognition theory. In *Situated cognition : social, semiotic, and psychological perspectives*, pages 57–70. 1997. ISBN 0805820388 (pbk.: alk. paper); 080582037X (hardcover: alk. paper). 167
- [535] G. M. Walton and G. L. Cohen. A question of belonging: race, social fit, and achievement. *Journal of personality and social psychology*, 92(1):82–96, 2007. ISSN 0022-3514. doi: 10.1037/0022-3514.92.1.82. 34
- [536] H. Wang and C.-T. Sun. Game reward systems: Gaming experiences and social meanings. In *DiGRA Conference*, 2011. 187
- [537] J. K. Waters. THE Journal, 2014. URL <https://thejournal.com/Articles/2014/05/14/Adaptive-Learning-Are-We-There-Yet.aspx?p=1>. 44
- [538] M. Watts. *Avatar Self-Identification, Self-Esteem, and Perceived Social Capital in the Real World: A Study of World of Warcraft Players and their Avatars*. University of South Florida, 2016. 189
- [539] A. Waytz, J. Heafner, and N. Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52:113–117, 2014. ISSN 00221031. doi: 10.1016/j.jesp.2014.01.005. 95
- [540] D. Weintrop and U. Wilensky. Program-to-play video games: Developing computational literacy through gameplay . pages 1–7, 2014. URL <http://ccl.northwestern.edu/papers/2014/GLS-2014final.pdf{%}5Cnpapers2://publication/uuid/13559C2B-81A3-474F-B5B9-C855BCD859AA>. 45

- [541] D. Weintrop, E. Beheshti, M. Horn, K. Orton, K. Jona, L. Trouille, and U. Wilensky. Defining Computational Thinking for Mathematics and Science Classrooms. *Journal of Science Education and Technology*, 25(1):127–147, 2016. ISSN 15731839. doi: 10.1007/s10956-015-9581-5. [43](#)
- [542] I. Wender. Relation of Technology, Science, Self-Concept, Interest, and Gender. *Journal of Technology Studies*, 30(3):43–51, 2004.
- [543] E. Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, 1998. ISBN 9780521430173. doi: 10.1016/j.jchas.2013.03.426. [167](#)
- [544] K. Werbach and D. Hunter. *For the win: How game thinking can revolutionize your business*. Wharton Digital Press, 2012. [187](#)
- [545] M. Westrom and A. Shaban. Intrinsic motivation in microcomputer games. *Journal of Research on Computing in Education*, 24(4):433, 1992. ISSN 08886504. doi: 10.1080/08886504.1992.10782018. URL <http://gateway.library.qut.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true{%&}db=afh{%&}AN=9609221618{%&}site=ehost-live>. [173](#)
- [546] L. S.-m. Whang and G. Chang. Lifestyles of virtual world residents: living in the on-line game "lineage". *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 7(5):592–600, 2004. ISSN 1094-9313. doi: 10.1109/CYBER.2003.1253430. [31](#), [131](#), [132](#), [136](#), [189](#)
- [547] R. W. White. Motivation reconsidered: The concept of competence. *Psychological review*, 66(5):297, 1959. [78](#), [197](#)
- [548] E. N. Wiebe, L. Williams, K. Yang, and C. Miller. Computer Science Attitude Survey. *NCSU CSC TR-2003-1*, 53(9):1689–1699, 2003. ISSN 1098-6596. doi: 10.1017/CBO9781107415324.004.

- [549] U. Wilensky. *Abstract meditations on the concrete and concrete implications for mathematics education*. 1991. URL <https://ccl.northwestern.edu/papers/concrete/>. 3
- [550] U. Wilensky. NetLogo. 1999. 45
- [551] U. Wilensky and I. Evanston. NetLogo: Center for connected learning and computer-based modeling. *Northwestern University, Evanston, IL*, pages 49–52, 1999. 38, 184
- [552] K. M. Williamson, L. Land, B. Butler, and H. B. Ndahi. A structured framework for using games to teach mathematics and science in K-12 classrooms. *The Technology Teacher*, 2004. 184, 188
- [553] K. Willsher, J. Penman, and Others. Engaging with schools and increasing primary school students' interest in science: An intersectoral collaboration. *Education in Rural Australia*, 21(2):87, 2011.
- [554] K. a. Wilson, W. L. Bedwell, E. H. Lazzara, E. Salas, C. S. Burke, J. L. Estock, K. L. Orvis, and C. Conkey. Relationships Between Game Attributes and Learning Outcomes: Review and Research Proposals. *Simulation & Gaming*, 40(2):217–266, 2009. ISSN 1046-8781. doi: 10.1177/1046878108321866. 26, 94, 166, 167, 173
- [555] J. M. Wing. Computational thinking. *Communications of the ACM*, 49(3):33–35, 2006. 40, 42, 43
- [556] J. B. Wise, A. E. Posner, and G. L. Walker. Verbal messages strengthen bench press efficacy. *Journal of Strength & Conditioning Research*, 18(1):26–29, 2004. ISSN 1064-8011. 155, 164
- [557] D. Wolber. App inventor and real-world motivation. In *Proceedings of the 42nd ACM technical symposium on Computer science education*, pages 601–606. ACM, 2011. 45

- [558] J. Wolfendale. My avatar, my self: Virtual harm and attachment. *Ethics and Information Technology*, 9(2):111–119, 2007. ISSN 1388-1957. doi: 10.1007/s10676-006-9125-z. URL <http://link.springer.com/10.1007/s10676-006-9125-z>. 31, 131, 132, 136, 189
- [559] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining, 2008. ISSN 0219-1377. 102
- [560] N. Yee and J. Bailenson. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Comm. Research*, jul 2007. ISSN 0360-3989. doi: 10.1111/j.1468-2958.2007.00299.x. 31, 44, 189, 214
- [561] N. Yee and J. Bailenson. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Communication Research*, 33(3):271–290, jul 2007. ISSN 0360-3989. doi: 10.1111/j.1468-2958.2007.00299.x. URL <http://doi.wiley.com/10.1111/j.1468-2958.2007.00299.x>. 132
- [562] N. Yee, N. Ducheneaut, M. Yao, and L. Nelson. Do Men Heal More When in Drag? *CHI 2011*, pages 1–4, 2011. 189, 214
- [563] M. F. Young. Instructional design for situated learning. *Educational Technology Research and Development*, 41(1):43–58, 1993. ISSN 10421629. doi: 10.1007/BF02297091. 167
- [564] L. Yuan and S. Powell. MOOCs and Open Education: Implications for Higher Education. *Cetis*, page 19, 2013. ISSN 1887-1542. doi: <http://publications.cetis.ac.uk/2013/667>. URL <http://publications.cetis.ac.uk/2013/667>. 19, 167
- [565] R. B. Zajonc. Emotions. In *The handbook of social psychology*, pages 591–632. 1998. 147
- [566] A. L. Zeldin and F. Pajares. Against the odds: Self-efficacy beliefs of women in

- mathematical, scientific, and technological careers. *American Educational Research Journal*, 37(1):215–246, 2000. 77
- [567] T. Zhang and B. Han. Experience reverses the red effect among Chinese stockbrokers. *PloS one*, 9(2):e89193, jan 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0089193. URL <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089193>. 148
- [568] R. Zheng, B. Stucky, M. McAlack, M. Menchana, and S. Stoddart. WebQuest learning as perceived by higher-education learners. *TechTrends*, 49(4):41–49, 2005. 141
- [569] Z.-J. Zhong and M. Z. Yao. Gaming motivations, avatar-self identification and symptoms of online game addiction. *Asian Journal of Communication*, 23(5):555–573, 2013. 189
- [570] G. Zicherman and C. Cunningham. Gamification by design. *Sebastopol, CA: O’Reilly Media*, 2011. 187
- [571] E. Zimmerman. Let the Games Be Games: Aesthetics, Instrumentalization & Game Design. In *Presentation at Game Developers Conference*, 2011. 167