Analysis of the Psychological and Production Effects of the Use of Gamification for Manufacturing Assembly

Makenzie Dolly
West Virginia University, mk0004@mix.wvu.edu

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Part of the Industrial Engineering Commons

Recommended Citation

This Dissertation is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Dissertation in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself. This Dissertation has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.
Analysis of the Psychological and Production Effects of the Use of Gamification for Manufacturing Assembly

Makenzie Dolly

Dissertation submitted to the Benjamin M. Statler College of Engineering and Mineral Resources at West Virginia University in partial fulfillment of the requirements for the degree of
PhD in
Industrial Engineering

Thorsten Wuest, Ph.D., Chair
Ashish Nimbarte, Ph.D.
Zhichao Liu, Ph.D.
Jannicke Baalsrud Hauge, Ph.D.
Mélanie Despeisse, Ph.D.

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia
2023

Keywords: gamification, smart manufacturing, operations, production, game elements, Industry 4.0

Copyright 2023 Makenzie Dolly
ABSTRACT

Analysis of the Psychological and Production Effects of the Use of Gamification for Manufacturing Assembly

Makenzie Dolly

In this dissertation, the applications of gamification for manufacturing with a focus on effects to workers and productivity were studied. Gamification is a relatively new research area, with the term being officially defined in 2010. Since then, several fields (education, health, and marketing) have benefitted from its application. Despite exhibiting strong potential, the application of gamification had remained rather unexplored in the manufacturing domain. To explore this further, by employing a comprehensive literature review, four research gaps were identified: the need for i) the use and acceptance of Deterding’s definition of gamification, ii) a clearer definition for various game element terms, iii) additional empirical research, and iv) the development of step-by-step guidelines for implementing gamification for manufacturing. The importance of Deterding’s definition was established through documentation and explanation in the state of the art. A classification framework was developed by sorting the game elements into eight groups based on characteristics of implementation specific to manufacturing. An empirical study focusing on implementing gamification for a monotonous assembly task was designed and completed. The experimental setup utilized a build kit for a Lego Telehandler (kit #42133). Cycle time data, Myers Briggs Type Indicators, and NASA TLX assessment data from 20 participants for 15 repetitive builds were collected, with data collection alone amassing well over 110 hours and approximately 1,100 data points. The results of the 98 unique analyses indicated that gamification had a significant effect on the productivity for the last build in a series of assembly tasks and temporal demand at the first build. While statistical significance was not found for many of the analyses, thorough discussions regarding trends in the data and limitations of the study indicate that, with additional research, statistical significance may potentially be established for these analyses as well.

To the best of our knowledge, this is the first study that systematically evaluates gamification as a possible manufacturing productivity improvement tool by considering manufacturing data (cycle time), task load of the user, and user-specific attributes (personality). The study makes the following contributions to the body of knowledge: i) supports the slowly growing set of empirical studies in gamification for manufacturing, ii) provides a transparent methodology that can be used by others to continue contributing empirical data on gamification for manufacturing, iii) showcases a method of design and implementation of gamification for manufacturing, and finally, iv) provides considerations for the future of research in gamification for manufacturing assembly tasks.
Table of Contents

Dedication .................................................................................................................................................. v

Acknowledgements ........................................................................................................................................ vi

List of Figures ............................................................................................................................................... vii

List of Tables .............................................................................................................................................. x

List of Abbreviations ................................................................................................................................ xii

List of Nomenclature ................................................................................................................................ xi

1. Introduction ........................................................................................................................................... 1

2. State of the Art ....................................................................................................................................... 3

2.1. Introduction ........................................................................................................................................ 3

2.2. Methodology ..................................................................................................................................... 4

2.3. Results and Discussions .................................................................................................................. 5

2.3.1. GfM Growth over the Years ........................................................................................................... 5

2.3.2. Current Stage of GfM Research ...................................................................................................... 6

2.3.3. Industry Sectors and Work Processes ........................................................................................... 8

2.3.4. Definitions of Gamification .......................................................................................................... 11

2.3.5. Game Design Elements .................................................................................................................. 13

2.3.6. Referring to “Game Elements” ...................................................................................................... 16

2.3.7. Technology Pairings ....................................................................................................................... 17

2.3.8. Benefits ......................................................................................................................................... 21

2.3.9. Limitations and Challenges .......................................................................................................... 22

2.3.10. Future Research Recommendations .......................................................................................... 24

2.4. Conclusions ....................................................................................................................................... 25

2.5. Research Problem ............................................................................................................................. 27

2.6. Dissertation Research Overview and Hypotheses ........................................................................... 27

3. Background ......................................................................................................................................... 29

3.1. Aspects of the Gamified Scenario ...................................................................................................... 29

3.1.1. Game Elements .............................................................................................................................. 29

3.1.2. Lego Set ......................................................................................................................................... 35

3.2. Variables of Interest for Data Collection ........................................................................................ 36

3.2.1. Cycle Time ...................................................................................................................................... 36

3.2.2. NASA Task Load Index ............................................................................................................... 36

3.2.3. Personality .................................................................................................................................... 37

4. Methodology ......................................................................................................................................... 38

4.1. Preparation and Development ......................................................................................................... 38

4.2. Overview of the Experiment ............................................................................................................. 42

4.3. Recruitment ....................................................................................................................................... 42

4.4. Participant Selection .......................................................................................................................... 44

4.5. Data Collection .................................................................................................................................. 45

4.6. Exit Survey ....................................................................................................................................... 50

5. Results and Discussions ....................................................................................................................... 52

5.1. Assumptions ....................................................................................................................................... 52

5.1.1. Outliers .......................................................................................................................................... 52

5.1.2. Standard Deviations ....................................................................................................................... 53

5.2. Hypothesis 1 ...................................................................................................................................... 55

5.3. Hypothesis 2 ...................................................................................................................................... 58

5.3.1. Difference between First and Last Cycle Time ........................................................................... 59
Dedication

To my husband.
Acknowledgements

First and foremost, thank you to my wonderful and amazing committee members, Zhichao Liu, Ph.D. (WVU), Mélanie Despeisse Ph.D. (Chalmers), Jannicke Baalsrud-Hauge, Ph.D. (KTH), Ashish Nimbarte, Ph.D. (WVU), and Thorsten Wuest, Ph.D. (WVU). Each and every one of you provided such fruitful discussions and insights throughout the journey. I truly enjoyed the final defense discussion, and I never thought that would be possible.

Thank you to all of the wonderful academics I’ve met and worked with over the years. Your willingness to support a young Ph.D. student and your friendliness was and is still so appreciated. I was so blessed to get to know so many kind and intelligent individuals.

Thank you to my advisor and mentor, Dr. Wuest. I’ve told many people over the years that, without you as my advisor, I probably wouldn’t have made it this far in the program. You have been an amazing mentor, in academia and research, and also in life. Your support and understanding on my tougher days meant more to me than I can express.

Thank you to my high school physics teacher, Don Walz. You exemplified the type of teacher I aspire to be (like using flying pigs and swinging bowling balls to teach physics concepts) and taught me to work for every ounce of what I want (because an 89.9% is not an A).

Thank you to my cat, Tigger, for catching some of my tears, but mostly for being a good desk decoration. Thank you to my dog, Maci, for giving me an excuse to step away from my desk for some play time and for bringing so much joy to every day.

Thank you to my best friend, Anna Billey, for coming into my life when I needed you most. From surviving class, planting flowers, planning my wedding, countless sleepovers, Writing Wednesdays, and endless commiserating, I couldn’t have done it without you.

Thank you to my forever best friend, Catherine “Piglet” Marrone, for being my biggest cheerleader even from way too far away.

Thank you to my husband, Byron. There aren’t enough words to express how much your undying and unwavering support has meant over the years. From encouraging me to pursue my dreams when we graduated undergrad (also thank you to Josh Bintrim for this one), to showing me tough love on the days I needed it most, to drawing festive signs of encouragement, to taking care of the house and the animals when I was on campus for 12+ hours a day for months at a time, to tolerating (and loving) me through my stress-induced spells of craziness, to being the first person to hug me when I passed my final defense, I couldn’t have asked for a better supporter. Thank you.

Thank you to my parents for raising me to be the hardworking, stubborn, strong-willed woman that I am. You’ve always supported every decision I’ve made on my own, except that one time I thought about quitting, and you talked me out of it; thank you. We might’ve argued when I was a teenager because I thought I “knew it all,” but now I have a title that says I actually do (totally kidding).

I could probably double the dissertation with the amount of thank you’s I have to express, but the internet told me I should keep it to one page. So, with that, thank YOU. Thank you to every friend, family member, acquaintance, student, peer, lab mate, supervisor, neighbor, teacher, mentor, reader, etc. over the years. I’m truly grateful for each and every one of you.
List of Figures

Figure 1. Overview of the structure of the dissertation. ................................................................. 2
Figure 2. Flow chart depicting the collection of papers for the literature review........................... 4
Figure 3. Growth of GfM publications (No. of papers per year (2012-2021)) ............................... 5
Figure 4. Industry sectors vs. work processes. ................................................................................. 10
Figure 5. Word cloud emphasizing game design elements based on recurrence in the lit rev. ....... 14
Figure 6. Pyramid of dynamics, mechanics, and components. ......................................................... 17
Figure 7. Classification framework depicting the eight groups for game elements. ......................... 32
Figure 8. Packaging for Lego Technic kit #42133, Telehandler used in this research. .................... 35
Figure 9. An example screen of the progress bar, shown while a participant is working on step 16. ... 39
Figure 10. An example screen of the progress bar, participant now working on step 17. ................. 39
Figure 11. “Your Badges” screen of the interface at the very start of the study. ............................... 40
Figure 12. An updated “Your Badges” screen after completion of the first build. ......................... 40
Figure 13. Finite badges awarded to participants. .......................................................................... 41
Figure 14. Infinite badge awarded to participants. ......................................................................... 41
Figure 15. Overview of the methodology for the experiment. ......................................................... 42
Figure 16. Sign posted around campus to recruit participants. ....................................................... 43
Figure 17. Screenshot of mass email sent campus wide to recruit participants. ......................... 44
Figure 18. Screenshot of "The Personality Types" app used to assess participant's MBTI. ............... 46
Figure 19. Disassembled Lego kit and build manual, provided to participants at the start of build. ... 46
Figure 20. View of materials provided to participant in gamified group. ....................................... 48
Figure 21. Build process for control participants. .......................................................................... 49
Figure 22. Build process for gamification participants. ................................................................. 49
Figure 23. Screenshots of the NASA TLX app used in the study. .................................................. 50
Figure 24. Cycle time 1 (CT1) in seconds, by treatment group with indication of outlier. ............... 53
Figure 25. Cycle time 1 (CT1) in seconds, by treatment group, for analyzing differences in var. .... 54
Figure 26. Dif_CT1.CT15 in seconds by treatment group, for analyzing differences in var. .......... 55
Figure 27. Histogram of cycle time 1 (CT1) in seconds for all participants. .................................... 56
Figure 28. Strip chart of CT1 in seconds by treatment group, for identifying a difference in means. ... 56
Figure 29. Histogram of average cycle time during visit 1 (AvgCT_V1) in seconds. ....................... 57
Figure 30. Strip chart of average cycle time from visit 1 (AvgCT_V1) in seconds by treatment group .... 58
Figure 31. Histogram of difference from cycle time 1 to cycle time 15 (Dif_CT1.CT15) in seconds. .. 59
Figure 32. Strip chart of Dif_CT1.CT15 by treatment group, for identifying a difference in means. ... 60
Figure 33. Histogram of Dif_AvgCTV1.AvgCTV5 in seconds. ....................................................... 61
Figure 34. Strip chart depicting Dif_AvgCTV1.AvgCTV5 in seconds by treatment group. ............... 61
Figure 35. Histogram of final attempt cycle times (CT15) in seconds. ......................................... 62
Figure 36. Strip chart last attempt cycle time (CT15) in seconds by treatment group. ................... 62
Figure 37. Histogram of the average cycle time of the last visit (AvgCT_V5) in seconds. ............... 64
Figure 38. Strip chart depicting AvgCT_V5 in seconds by treatment group. .................................. 64
Figure 39. Histogram of NASA TLX Mental Demand scores from visit 1 (TLX.MD.V1). ............... 67
Figure 40. Strip chart depicting TLX.MD.V1 by treatment group. ................................................. 67
Figure 41. Histogram of NASA TLX Physical Demand scores from visit 1 (TLX.PD.V1). ............... 68
Figure 42. Strip chart depicting TLX.PD.V1 by treatment group. ................................................. 69
Figure 43. Histogram of NASA TLX Temporal Demand scores from visit 1 (TLX.TD.V1). .......... 70
Figure 44. Strip chart depicting TLX.TD.V1 by treatment group ................................................................. 70
Figure 45. Histogram of NASA TLX Performance scores from visit 1 (TLX.P.V1). .................................... 71
Figure 46. Strip chart depicting TLX.P.V1 by treatment group ................................................................. 72
Figure 47. Histogram of NASA TLX Effort scores from visit 1 (TLX.E.V1). ............................................. 73
Figure 48. Strip chart depicting TLX.E.V1 by treatment group ................................................................. 73
Figure 49. Histogram of NASA TLX Frustration scores from visit 1 (TLX.F.V1) ..................................... 75
Figure 50. Strip chart depicting TLX.F.V1 by treatment group ................................................................. 75
Figure 51. Histogram of Dif_MDV1.MDV5 .............................................................................................. 77
Figure 52. Strip chart depicting Dif_MDV1.MDV5 by treatment group .................................................... 78
Figure 53. Histogram of Dif_PDV1.PDV5 ............................................................................................... 79
Figure 54. Strip chart depicting Dif_PDV1.PDV5 by treatment group .................................................... 80
Figure 55. Histogram of Dif_TDV1.TDV5 ............................................................................................... 81
Figure 56. Strip chart depicting Dif_TDV1.TDV5 by treatment group .................................................... 81
Figure 57. Histogram of Dif_PV1.PV5 ...................................................................................................... 82
Figure 58. Strip chart depicting Dif_PV1.PV5 by treatment group ............................................................ 83
Figure 59. Histogram of Dif_EV1.EV5 ..................................................................................................... 84
Figure 60. Strip chart Dif_EV1.EV5 by treatment group ........................................................................... 84
Figure 61. Histogram of Dif_FV1.FV5 ..................................................................................................... 85
Figure 62. Strip chart depicting Dif_FV1.FV5 by treatment group ............................................................ 86
Figure 63. Histogram of NASA TLX mental demand scores from visit five (TLX.MD.V5) ..................... 88
Figure 64. Strip chart depicting TLX.MD.V5 by treatment group ............................................................ 88
Figure 65. Histogram of NASA TLX physical demand scores from visit five (TLX.PD.V5) ..................... 89
Figure 66. Strip chart depicting TLX.PD.V5 by treatment group ............................................................. 90
Figure 67. Histogram of NASA TLX temporal demand scores from visit five (TLX.TD.V5) ................... 91
Figure 68. Strip chart depicting TLX.TD.V5 by treatment group ............................................................. 91
Figure 69. Histogram of NASA TLX performance scores from visit five (TLX.P.V5) ............................... 92
Figure 70. Strip chart depicting TLX.P.V5 by treatment group ................................................................. 93
Figure 71. Histogram of NASA TLX effort scores from visit five (TLX.E.V5) ......................................... 94
Figure 72. Strip chart depicting TLX.E.V5 by treatment group ................................................................. 95
Figure 73. Histogram of NASA TLX frustration scores from visit five (TLX.F.V5) ............................... 96
Figure 74. Strip chart depicting TLX.F.V5 by treatment group ................................................................. 96
Figure 75. Histogram of average NASA TLX mental demand scores (AvgMD) ....................................... 98
Figure 76. Strip chart depicting AvgMD by treatment group ................................................................. 99
Figure 77. Histogram of average NASA TLX physical demand scores (AvgPD) ...................................... 100
Figure 78. Strip chart depicting AvgPD by treatment group ................................................................. 100
Figure 79. Histogram of average NASA TLX temporal demand scores (AvgTD) ................................. 101
Figure 80. Strip chart depicting AvgTD by treatment group ................................................................. 102
Figure 81. Histogram of average NASA TLX performance scores (AvgP) ............................................ 103
Figure 82. Strip chart depicting average NASA TLX performance scores (AvgP) by treatment group .... 103
Figure 83. Histogram of average NASA TLX effort scores (AvgE) .......................................................... 104
Figure 84. Strip chart depicting average NASA TLX effort scores (AvgE) by treatment group ............ 105
Figure 85. Histogram of average NASA TLX frustration scores (AvgF) ................................................. 106
Figure 86. Strip chart depicting average NASA TLX frustration scores (AvgF) by treatment group ....... 106
Figure 87. Strip chart depicting TLX.PD.V1 by introvert/extrovert personality ..................................... 108
Figure 88. Strip chart depicting AvgP by introvert/extrovert personality .................................................. 109
Figure 89. Strip chart depicting TLX.P.V1 by knowing/sensing personality ............................................. 110
Figure 90. Strip chart depicting AvgP by knowing/sensing personality .................................................... 111
Figure 91. Scatter plot of AvgTD by self-competitiveness with line of relation of the two variables ....... 113
Figure 92. Scatter plot of AvgP by self-competitiveness with line of relation of the two variables .......... 114
Figure 93. Scatter plot of AvgF by self-competitiveness with line of relation of the two variables ............ 115
Figure 94. Scatter plot of AvgPD by patience with line of relation of the two variables ...................... 117
Figure 95. Scatter plot of AvgF by patience with line of relation of the two variables ......................... 118
Figure 96. Strip chart depicting cycle time 1 (CT1) in seconds by student status ...................................... 121
Figure 97. Strip chart depicting Dif_CT1.CT15 in seconds by student status .......................................... 121
Figure 98. Strip chart depicting cycle time 15 (CT15) in seconds by student status ................................. 122
Figure 99. Strip chart depicting cycle time 1 (CT1) in seconds by self-rated experience ..................... 122
Figure 100. Strip chart depicting Dif_CT1.CT15 in seconds by self-rated experience .............................. 123
Figure 101. Strip chart depicting cycle time 15 (CT15) in seconds by self-rated experience ................. 123
Figure 102. Strip chart depicting cycle time 1 (CT1) in seconds by age .............................................. 124
Figure 103. Strip chart depicting Dif_CT1.CT15 in seconds by age ...................................................... 124
Figure 104. Strip chart depicting cycle time 15 (CT15) in seconds by self-rated experience ............... 125
Figure 105. Strip chart depicting accuracy pressure by treatment group .............................................. 126
Figure 106. Strip chart depicting time pressure by treatment group ................................................... 126
Figure 107. Strip chart depicting expected performance in the “other” group by treatment group ....... 127
Figure 108. Line graphs depicting each cycle time for each participant throughout the study ............. 147
List of Tables
Table 1. Tabulation of results for the categorization of papers for different types of research ................. 7
Table 2. Industry sectors present in the literature review. ......................................................................... 9
Table 3. Work processes present in literature review ............................................................................... 9
Table 4. Tabulation of key definitions in gamification research.............................................................. 12
Table 5. Description and implications of the top six commonly referenced game design elements ...... 15
Table 6. Alternative terms to refer to "game elements." ....................................................................... 16
Table 7. Technology pairings .................................................................................................................. 18
Table 8. Technologies presently used alongside GfM and their descriptions and applications .............. 20
Table 9. Benefits of GfM as described in the literature, ranked ............................................................ 21
Table 10. Benefits of GfM in two large categories: psychological and production .................................. 22
Table 11. Limitations and challenges of using GfM, according to the literature .................................... 23
Table 12. Future research recommendations to further the field of GfM, according to the literature ... 25
Table 13. Game elements classified into their respective group. ............................................................. 33
Table 14. Sample size of experiments as seen from articles considered in the State of the Art ............... 45
Table 15. Standard deviation for two different variables, to compare the equality of var. ................. 54
Table 16. Treatment group means, and resulting p-value of Welch's t-test for H1 variables .................. 58
Table 17. Treatment group means, and resulting p-value of Welch's t-test for H2 variables .......... 65
Table 18. Treatment group means, and resulting p-value of Welch's t-test for H3 variables ............ 76
Table 19. Treatment group means, and resulting p-value of Welch's t-test for Dif_TLXV1.TLXV5 ........ 87
Table 20. Treatment group means, and resulting p-value of Welch's t-test for TLX.V5 ..................... 97
Table 21. Treatment group means, and resulting p-value of Welch's t-test for AvgTLX variables .......... 107
Table 22. Treatment group means, and resulting p-value of Welch's t-test for MBTI vs. AvgTLX .......... 111
Table 23. Correlation values for competitive nature and NASA TLX scores ........................................ 116
Table 24. Correlation values for patience and NASA TLX scores ....................................................... 119
Table 25. Resulting p-value of Welch's t-test for variables related to productivity vs. MBTI ............ 120
Table 26. Count of participants in each treatment group ................................................................. 128
Table 27. Count of participants in each group of the Introvert/Extrovert MBTI personality pairing ... 128
Table 28. Count of participants in each group of the Knowing/Sensing MBTI personality pairing ... 128
Table 29. Count of participants in each group of the Thinking/Feeling MBTI personality pairing ...... 129
Table 30. Count of participants in each group of the Judging/Perceiving MBTI personality pairing ... 129
Table 31. Count of participants in each group of student status ....................................................... 129
Table 32. Count of participants in each group of self-rated experience level .................................... 129
Table 33. Descriptive statistics for every variable included in the study ......................................... 129
Table 34. Condensed list of conclusions for each of the five hypotheses (H1-5) of the study ........... 137
Table 35. Count of papers, by overlapping industries and work processes within reviewed papers ... 146
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>Augmented reality</td>
</tr>
<tr>
<td>E</td>
<td>Effort</td>
</tr>
<tr>
<td>F</td>
<td>Frustration</td>
</tr>
<tr>
<td>GfM</td>
<td>Gamification for manufacturing</td>
</tr>
<tr>
<td>H1</td>
<td>Hypothesis 1</td>
</tr>
<tr>
<td>H2</td>
<td>Hypothesis 2</td>
</tr>
<tr>
<td>H3</td>
<td>Hypothesis 3</td>
</tr>
<tr>
<td>H4</td>
<td>Hypothesis 4</td>
</tr>
<tr>
<td>H5</td>
<td>Hypothesis 5</td>
</tr>
<tr>
<td>ID</td>
<td>Identification</td>
</tr>
<tr>
<td>IRB</td>
<td>Institutional Review Board</td>
</tr>
<tr>
<td>IVA</td>
<td>Intelligent virtual assistant</td>
</tr>
<tr>
<td>KPI</td>
<td>Key performance indicator</td>
</tr>
<tr>
<td>MD</td>
<td>Mental demand</td>
</tr>
<tr>
<td>MS</td>
<td>Microsoft</td>
</tr>
<tr>
<td>NASA</td>
<td>National Aeronautics and Space Administration</td>
</tr>
<tr>
<td>P</td>
<td>Performance</td>
</tr>
<tr>
<td>PD</td>
<td>Physical demand</td>
</tr>
<tr>
<td>s</td>
<td>Seconds</td>
</tr>
<tr>
<td>TD</td>
<td>Temporal demand</td>
</tr>
<tr>
<td>TLX</td>
<td>Task load index</td>
</tr>
<tr>
<td>VR</td>
<td>Virtual reality</td>
</tr>
</tbody>
</table>
List of Nomenclature
In order of appearance in 5. Results and Discussions.

<table>
<thead>
<tr>
<th>Nomenclature</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT1</td>
<td>First attempt cycle time</td>
</tr>
<tr>
<td>CT15</td>
<td>Last attempt cycle time</td>
</tr>
<tr>
<td>AvgCT_V1</td>
<td>First visit average cycle time</td>
</tr>
<tr>
<td>AvgCT_V5</td>
<td>Last visit average cycle time</td>
</tr>
<tr>
<td>Dif_CT1.CT15</td>
<td>Difference between first and last attempt cycle time</td>
</tr>
<tr>
<td>Dif_AvgCT1.AvgCTV5</td>
<td>Difference between first and last visit average cycle time</td>
</tr>
<tr>
<td>TLX.MD.V1</td>
<td>Task load index mental demand score at first visit</td>
</tr>
<tr>
<td>TLX.PD.V1</td>
<td>Task load index physical demand score at first visit</td>
</tr>
<tr>
<td>TLX.TD.V1</td>
<td>Task load index temporal demand score at first visit</td>
</tr>
<tr>
<td>TLX.PV1</td>
<td>Task load index performance score at first visit</td>
</tr>
<tr>
<td>TLX.E.V1</td>
<td>Task load index effort score at first visit</td>
</tr>
<tr>
<td>TLX.F.V1</td>
<td>Task load index frustration score at first visit</td>
</tr>
<tr>
<td>Dif_MDV1.MDV5</td>
<td>Difference between first and last visit mental demand scores</td>
</tr>
<tr>
<td>Dif_PDV1.PDV5</td>
<td>Difference between first and last visit physical demand scores</td>
</tr>
<tr>
<td>Dif_TDV1.TDV5</td>
<td>Difference between first and last visit temporal demand scores</td>
</tr>
<tr>
<td>Dif_PV1.PV5</td>
<td>Difference between first and last visit performance scores</td>
</tr>
<tr>
<td>Dif_EV1.EV5</td>
<td>Difference between first and last visit effort scores</td>
</tr>
<tr>
<td>Dif_FV1.FV5</td>
<td>Difference between first and last visit frustration scores</td>
</tr>
<tr>
<td>TLX.MD.V5</td>
<td>Task load index mental demand score at fifth visit</td>
</tr>
<tr>
<td>TLX.PD.V5</td>
<td>Task load index physical demand score at fifth visit</td>
</tr>
<tr>
<td>TLX.TD.V5</td>
<td>Task load index temporal demand score at fifth visit</td>
</tr>
<tr>
<td>TLX.PV5</td>
<td>Task load index performance score at fifth visit</td>
</tr>
<tr>
<td>TLX.E.V5</td>
<td>Task load index effort score at fifth visit</td>
</tr>
<tr>
<td>TLX.F.V5</td>
<td>Task load index frustration score at fifth visit</td>
</tr>
<tr>
<td>AvgMD</td>
<td>Average mental demand score across all five visits</td>
</tr>
<tr>
<td>AvgPD</td>
<td>Average physical demand score across all five visits</td>
</tr>
<tr>
<td>AvgTD</td>
<td>Average temporal demand score across all five visits</td>
</tr>
<tr>
<td>AvgP</td>
<td>Average performance score across all five visits</td>
</tr>
<tr>
<td>AvgE</td>
<td>Average effort score across all five visits</td>
</tr>
<tr>
<td>AvgF</td>
<td>Average frustration score across all five visits</td>
</tr>
</tbody>
</table>
1. Introduction

Please note: significant portions of this section have been previously published in my peer-reviewed article, titled “Current State of Research & Outlook of Gamification for Manufacturing”, published as the first GfM paper in the premier Journal of Manufacturing Systems [IF: 9.498]. (Keepers et al., 2022)

Gamification, the application of game design elements to non-game contexts (Deterding et al., 2011), is making its way into nearly everyone’s lives through shopping rewards programs, fitness tracking challenges, social media platforms, and incentive programs at work. The premise of gamification aligns with the human nature to compete and can incite positive change in various forms, such as, motivating users, initiating learning, and enhancing cooperation (Lithoxoidou et al., 2017). Most commonly, gamification can be seen as a tool to promote motivation and to encourage users to complete a task better than they otherwise would have (Bahr et al., 2021). In some cases, this can mean completing the task faster, more efficiently, and/or with fewer mistakes such as in an assembly task, or in the case of fitness tracking, simply completing the task.

Given current applications of gamification in other fields, such as retail, marketing, and healthcare, gamification may prove to be a promising technique within the field of manufacturing. While examples of gamification can easily be pointed out in everyday life, the examples of gamification implemented in an industrial setting, more specifically in manufacturing systems and facilities (i.e., the shop floor), are rare and often hard to find. Some theoretical applications of gamification for manufacturing include: instances of single-player approaches to incite healthy competition between industrial workers and their workstations to reduce cycle times and/or multi-player approaches to enable industrial workers to communicate and cooperate in joint efforts in a production line or shop floor such as collaborative problem solving to advance manufacturing systems. To bridge the gap from applications of gamification in non-manufacturing settings to the use of gamification in manufacturing, we focus on two key areas, i) the effects on production (i.e., productivity and throughput) and ii) the effects on workers (i.e., physical exertion, stress, and frustration).

With gamification providing benefits such as motivation, increased happiness, reduced stress, and improved productivity, it seems logical to apply gamification to the world of manufacturing. At surface level, these benefits are likely to provide manufacturing systems with many positive outcomes and seemingly few negative effects. Increased employee motivation can translate to increased production output and a focus on lowering the rate of accidents and waste, increased happiness and less stress can improve employee retention rates and reduce burnout, and improved productivity would directly impact the manufacturing company’s bottom-line. So, why is there not more buzz about gamification for manufacturing?

Please note: This line denotes the conclusion of the material that has been acquired from (Keepers et al., 2022). The following material is new to support the dissertation document.

This dissertation aims to contribute to the Gamification for Manufacturing (GfM) body of knowledge by providing new insights based on empirical data, concrete recommendations for the future research of GfM, as well as recommendations for design and implementation of gamification for industrial settings. We begin in 2. State of the Art by taking an in-depth look at GfM with a comprehensive critical review of the scientific literature to explore the current state and outlook of the field. From this literature review, research gaps are identified, and four recommendations are made to support the GfM research
community, as well as to inform the research herein. 3. Background describes the background of the various components used to support the development of the empirical study, including the classification of game elements and data collection. The methodology of the empirical study is described in thorough detail in 4. Methodology. Thorough analysis of results and detailed discussions are provided in 5. Results and Discussions. In 6. Holistic Discussion, we look at the entire collection of results and discussions across the five hypotheses and other results of interest to provide a holistic discussion regarding GfM. The dissertation culminates in 7. Conclusion by wrapping up all of the information included herein, as well as provides limitations of the study and future research recommendations (Figure 1).

Figure 1. Overview of the structure of the dissertation.
2. State of the Art

Please note: significant portions of this section have been acquired from my published, peer-reviewed article, titled “Current State of Research & Outlook of Gamification for Manufacturing” published as the first GfM paper in the premier Journal of Manufacturing Systems [IF: 9.498]. (Keepers et al., 2022)

This section presents the current state of research for and an outlook of Gamification for Manufacturing (GfM), particularly manufacturing operations in industrial settings. This in-depth analysis examines the trends in published (2021 and earlier) conceptual design works, case studies, experiments, and literature reviews that are shaping the future design work of gamification in the manufacturing sector. Gamification, as a general area of research, has been a fast-growing topic of interest for many years across various industries and academic research groups (Keepers et al., 2020). Although the benefits of gamification are likely to apply to the manufacturing sector, limited research has been conducted as it relates to GfM operations outside of academic settings. Additionally, there is a limited number of scientific publications available that address the body of research as a whole to understand the current state and outlook of GfM. This review of literature works to close this research gap by firstly synthesizing the body of knowledge of GfM, and secondly, based on the findings obtained, provides four concrete recommendations for the continuous exploration of gamification trends and opportunities for manufacturing operations.

2.1. Introduction

This subsection provides a comprehensive critical review of the scientific literature of GfM in an effort to explore the current state and outlook of the field. This is an extension of a previous work that we completed in 2020, titled: Gamification of Operational Tasks in Manufacturing (Keepers et al., 2020). This work gathers additional papers by expanding the search terms and including references up to the year 2021, and it provides a more comprehensive and critical review of each paper in comparison to the previous work. The goal of this critical review is to provide the manufacturing community, within industry and academia, with an overview of current information available about GfM, including its definition, benefits and limitations, active areas of application, game elements, and various other topics as they relate to this technique. Most importantly, as gamification is a new and developing topic for the manufacturing sector, this section provides four concrete recommendations for the next steps in advancing GfM through research and application based on the conducted analysis.

Compared to other literature reviews in the scientific body of knowledge for the topic of GfM, this work differs by providing mainly an industrial, manufacturing systems perspective, and by excluding the manufacturing education domain. Additionally, the most recently published in-depth literature review on the topic is from 2018 (Warmelink et al., 2018), so this critical review provides a much more up-to-date look on this fast-growing topic by including an additional four years of reviewed research. 2.2. Methodology covers the methodology used for this critical review, including descriptions of the selection process of the papers included in it. 2.3. Results and Discussions centers around the results and discussion of the current state of research on GfM, particularly on manufacturing operations. Lastly, 2.4. Conclusions details the key findings and conclusions of the state-of-the-art and practice of GfM, and details four recommendations for the continuous exploration of gamification trends and opportunities for manufacturing operations.
2.2. Methodology

This critical review was structured according to recommendations from (Webster & Watson, 2002). In this subsection, we expand the scope of their previous exploratory work (Keepers et al., 2020) through broader and more inclusive search terms yielding a more comprehensive set of papers. Thus, this literature review provides a better representation of the actual body of knowledge of GfM. With the previous work only considering “manufacturing,” the addition of “operations” (i.e., manufacturing operations), “production” (i.e., production/manufacturing systems), and “logistics” (i.e., intra-logistics) as keywords especially addresses previous limitations. Therefore, this critical review aims to understand and derive the current research scope of GfM as it relates to industry settings and practices within the manufacturing sector. More specifically, we explicitly defined subtopics to be excluded to ensure that the literature included in this critical review was relevant, specifically for operations that happen at the factory and manufacturing systems level. While this is not a comprehensive list of all potential application areas where GfM can be put into practice, these are areas that resulted from the search and were deemed to be unrelated to core manufacturing operations. The areas or subtopics which were considered out of scope include:

- Service operations (as opposed to manufacturing operations).
- Manufacturing educational settings (as opposed to manufacturing industrial settings).

Conversely, the following topics were considered in scope:

- Manufacturing systems.
- Production systems (excluding service systems).
- Manufacturing operations.
- Logistics (particularly intra-logistics as the logistical flow of material goods within the walls of a factory, including warehousing).
- On-the-job training.
- Professional development.

To begin the review, the following search of title, abstract, and keywords was used on the Scopus database: TITLE-ABS-KEY((gamif*) AND (operation* OR manufactur* OR production* OR logistic)). During the paper selection process, only papers that were available in English were included. Other limiting factors for the search included:

- Document type: conference paper, journal article, or book chapter.
- Subject areas: computer science, engineering, or business.

An outline of the search is provided in Figure 2. The search yielded 515 unique papers. Of these 515 papers, by comparing the pre-defined scope with the titles and abstracts in a first step, 65 papers were considered relevant and selected to be read in their entirety. Of these 65 papers, 35 papers were determined to be within scope for this literature review. These 35 papers were read in their entirety and notes were compiled to aid in thorough reviews. Using these notes and referencing back to the full article, conclusions were drawn and are discussed in the following sections.

Figure 2. Flow chart depicting the collection of papers for the literature review.
2.3. Results and Discussions
The results and discussions section covers all findings from the literature review, including but not limited to the growth of GfM over the years (i.e., from 2012 to 2021), the varying definitions for gamification in the manufacturing context, and the numerous benefits this technique offers – particularly for the manufacturing domain. For each topic, a brief introduction is provided, and then broken into results and discussions (i.e., the critical review). The results of each topic are considered objective and provide the direct findings from the literature review without interpretation by the authors to reduce bias. The discussions of each topic are subjective in the sense that they provide interpretations from the research team and work to describe what the objective findings may indicate for the gamification technique in a manufacturing context. This subsection of the critical review is a key contributor by providing the foundation for the recommendations in the following subsection as well as future research in this area.

2.3.1. GfM Growth over the Years
Gamification has found its way into many of the products and applications that humans use in their daily life and is often labeled as a “fast-growing topic” of research interest for academia and industry. To understand how quickly this topic is growing, we depicted the year of publication for the papers which were included in this literature review in Figure 3. Recalling from the methodology section, the selection of papers was strict, so the chart of growth only shows the growth of GfM applications. Considering the growth of gamification as a general topic would show a more significant and more obvious trend of increase (Keepers et al., 2020).

![Figure 3. Growth of GfM publications (No. of papers per year (2012-2021))](image)

2.3.1.1. GfM Growth – Results
From Figure 3, the overall trend between 2012 to 2021 shows an overall increase in the number of papers covering GfM over the last decade. However, this growth is not linear and appears to have local peaks in 2016 and 2020 and local valleys in 2013, 2014, and 2018.

It is important to note that the papers included in this literature review were compiled in March 2021. While it is likely that by this time the majority of the 2021 papers had been published and were available
on the Scopus database, there may be additional papers added after the data set was updated and finalized. This dataset is assumed to be complete and timely for the included years 2012-2021.

2.3.1.2. GfM Growth – Discussion
Reasons for the variability of publications between years are most likely attributable to the relatively small sample size of 35 papers. Before 2016, five of the six papers published shared the same first author, Professor Oliver Korn (Korn, 2012; Korn, Boffo, et al., 2015; Korn et al., 2012, 2014; Korn, Funk, et al., 2015). From reading these papers, it is indicative that these are connected to one or a select few similar research projects. Professor Korn later authored two more papers connected to his previous research work, one in 2016 and a final paper in 2017 (Korn et al., 2016, 2017). Following Professor Korn’s papers throughout the years, the progress and development of the underlying research project is promising and highlights the potential presented by the gamification technique in the manufacturing domain. This is highlighted in order to showcase that the research field of GfM was primarily dominated by only a few research teams and projects, especially in the early years.

Most excitingly, in 2020, the second most recent year included in the search, emerged as the year with the most publications, directly followed by 2021, 2019, and 2016 with the second most publications. This finding supports the initial statement that gamification is a fast-growing research field, including in the particular domain of interest of this critical review: “manufacturing”.

2.3.2. Current Stage of GfM Research
Beyond understanding the growth of gamification within the manufacturing domain itself, we were interested in dissecting the current state of research on GfM. To do this, we considered the type of publication as a first step. We wanted to better understand whether this research topic was still centered on the earlier stages of defining the problem, had transitioned already to the stage of developing final concrete results, or otherwise. This is an important aspect to consider as it provides higher-level insights into the lifespan and outlook of GfM research and what to expect from recently published work as well as the future research trajectory.

2.3.2.1. Current Stage of GfM Research – Results
From most exploratory to more concrete research, the authors believe that the type of publications would progress as follows:

1. Literature review – a review of already available publications wherein literature is synthesized to understand the current scope of the field.
2. Conceptual design work – often supported by a literature review, involves developing a framework, tool, set of guidelines, or similar to support the future design work of the topic.
3. Case study – a mostly qualitative review of a real case, often including generalized statements pointing towards a possible trend.
4. Experiment – often a quantitative review of a scientifically developed experiment, typically including statistical testing of hypotheses.

We classified each paper included in this literature review as relating to six different categories – the four mentioned above, a fifth one associated with case studies culminating in conceptual design work (a step beyond completing a case study), and a sixth one related to experiments culminating in conceptual design work (a step beyond completing an experiment) (see Table 1). Note that papers were classified for the most suitable category, with no paper assigned multiple different categories. The two additional
overlapping categories were for papers that had a near-equal split of attention between the case study/experiment and the conceptual design work. Furthermore, many papers begin with a brief literature review to support the introduction or background of their work. This would not support the paper being classified as a literature review. Literature reviews are papers that solely focus on revisiting the literary corpus.

Table 1. Tabulation of results for the categorization of papers for different types of research.

<table>
<thead>
<tr>
<th>Type of Research</th>
<th>No. of Papers</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual Design Work</td>
<td>12</td>
<td>(Elitok &amp; Bolat, 2021; Hellebrandt et al., 2019; Hense et al., 2014; Korn, 2012; Korn et al., 2012, 2017; Plakas et al., 2020; Reis et al., 2020; Schuldt &amp; Friedemann, 2017; Sochor et al., 2021; Ulmer et al., 2020a, 2020b)</td>
</tr>
<tr>
<td>Experiment</td>
<td>9</td>
<td>(Bräuer &amp; Mazarakis, 2019; Hellebrandt et al., 2020; Korn, Boffo, et al., 2015; Korn et al., 2014, 2016; Ohlig et al., 2021; Ponis et al., 2020; Roh et al., 2016; Seo et al., 2021)</td>
</tr>
<tr>
<td>Case Study</td>
<td>4</td>
<td>(Bahr et al., 2021; Lithoxoidou et al., 2017; Pereira et al., 2018; Putz et al., 2019)</td>
</tr>
<tr>
<td>Literature Review</td>
<td>4</td>
<td>(Alavesa et al., 2019; Keepers et al., 2020; Markopoulos et al., 2015; Warmelink et al., 2018)</td>
</tr>
<tr>
<td>Case Study → Conceptual Design Work</td>
<td>3</td>
<td>(Butean et al., 2019; Lee et al., 2016; Siemens Energy, Industrial Application, Turbo Machinery et al., 2021)</td>
</tr>
<tr>
<td>Experiment → Conceptual Design Work</td>
<td>3</td>
<td>(Korn, Funk, et al., 2015; Lessel et al., 2016; Liu et al., 2018)</td>
</tr>
</tbody>
</table>

The two most common types of research were “experimental” and “conceptual design work,” with a total of twelve papers implementing experimentation. Additionally, six papers worked to provide insights into GfM by supporting conceptual design work based on their case study or experiment results. The third most common type of research was “case studies.” The least common type of research was “literature reviews” with only four papers focusing entirely on reviewing the current body of knowledge.

2.3.2.2. Current Stage of GfM Research – Discussion

The limited quantity of literature reviews is not surprising for a few reasons. First, there is a limited amount of research available to be reviewed. Second, a significant increase in literature reviews would be unlikely to differentiate themselves from previous literature reviews, and third and lastly, these would not provide additional support to the GfM research community.

An additional reason as to why there may only be a few relevant literature reviews for this review paper is due to the selection criteria. As gamification began in the education sector, there are often many papers focusing on gamification in education for manufacturing, and thus many literature reviews would also
focus on these papers. Considering that this literature review specifically excluded the academic setting, this may have resulted in the exclusion of literature reviews which focused on manufacturing education.

A comparably high quantity of conceptual design work indicates that researchers are beginning to formulate their thoughts into possible actionable plans for further research on GfM. Additionally, researchers are developing guidelines and frameworks to support the next phase of research: case studies and experimentation. Considering the five papers which had a case study or experiment culminating in conceptual design work, these support the fact that researchers can develop concrete findings to support the further exploration and implementation of GfM. In essence, the quantity of conceptual design work compared to the entirety of the body of research for GfM indicates that this research field is building a solid foundation to support its future growth.

The use of case studies is comparably low to other types of research work. This finding is understandable and to be expected given the in-depth nature of implementing GfM. Full implementation of a gamified scenario can be highly-involved and require significant hands-on practice. Additionally, there are few case studies available at this time since gamification is not readily implemented in industrial settings yet.

Given the infancy of GfM, the large number of experiments is unexpected and warrants a closer look. This seems to be partially due to the existence of multiple publications for related experiments, for example, Korn et al. account for four of the twelve experiments and all four of those experiments were related to a similar project. Although surprising, the high quantity of experiments indicates good promise that researchers are excited and able to test hands-on applications of GfM. The majority of the experiments included in this literature review are off-the-shop-floor experiments, primarily taking place in an academic lab setting. While this is not ideal, this provides a safe area to test out gamification for the efficacy of use in a real production environment.

2.3.3. Industry Sectors and Work Processes

As research evolves, it often begins with a general explorative approach. Then, as researchers gain a basic understanding of the new and emerging research topic, they develop use cases in additional niche research areas. By reviewing the industry sectors and work processes that are being discussed in the published GfM research included in this literature review, it was possible to understand which industry sectors and work processes have already been evaluated.

The “industry sector” was determined by reading the papers and classifying the type of facility or product which was being discussed. Some examples of the “industry sector” include automotive, electronics, and warehousing. If a paper discussed multiple sectors or did not clearly state the type of facility or product, the paper was classified as general production or general logistics. General production papers focus on the manufacturing of goods, while general logistics papers focus on the movement or transportation of goods within the walls of a factory. If a paper did not fit specifically into any of the previous categories, it was placed into the “other, general” category.

For the “work processes,” the papers were classified based on the type of operational process (operation) being addressed. Some examples of “work processes” include bolt-tightening and CNC machining. When a paper was not clear on the specific work process (operation), it was classified as “general operations,” “general manufacturing,” or “general assembly.” To distinguish general operations, general assembly, and general manufacturing in the remainder of the paper, the following definitions were abided: General assembly refers to the assembly or joining of various parts or modules to create a (new) product or system.
General manufacturing includes but is not limited to primary and secondary manufacturing processes such as subtractive or additive, as well as other shop floor manufacturing activities that do not fall under assembly (Harik & Wuest, 2020). General operations include supporting processes and operations such as business processes and other research that did not classify as assembly or manufacturing. When a paper discussed multiple applications or was not clear which category it best belonged to, it was classified as “other, general.”

2.3.3.1. Industry Sectors and Work Processes – Results

The majority of papers were classified as general production (see Table 2). Of the more specific categories, electronics manufacturing, air transport, and equipment manufacturing had the least number of papers. However, not all of the specific categories were limited in their contribution, as the sheltered workplace and automotive manufacturing categories had the second and third highest number of papers at a count of five and four papers, respectively.

Table 2. Industry sectors present in the literature review.

<table>
<thead>
<tr>
<th>Industry</th>
<th>No. of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>General production</td>
<td>13</td>
</tr>
<tr>
<td>Sheltered workplace</td>
<td>5</td>
</tr>
<tr>
<td>Automotive manufacturing</td>
<td>4</td>
</tr>
<tr>
<td>Warehouse</td>
<td>3</td>
</tr>
<tr>
<td>General intralogistics</td>
<td>3</td>
</tr>
<tr>
<td>Other, general</td>
<td>3</td>
</tr>
<tr>
<td>Electronics manufacturing</td>
<td>2</td>
</tr>
<tr>
<td>Air transport</td>
<td>1</td>
</tr>
<tr>
<td>Equipment manufacturing</td>
<td>1</td>
</tr>
</tbody>
</table>

In the work processes breakdown, general assembly had the greatest number of papers, with a total of eleven papers (see Table 3). There is a significant drop in papers, down to five papers for the next highest category, which was order picking. Only two papers looked at CNC machine operations, an established manufacturing process, to be subjected to gamification.

Table 3. Work processes present in literature review.

<table>
<thead>
<tr>
<th>Work Process</th>
<th>No. of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>General assembly</td>
<td>11</td>
</tr>
<tr>
<td>Order picking</td>
<td>5</td>
</tr>
<tr>
<td>General manufacturing</td>
<td>5</td>
</tr>
<tr>
<td>Other, general</td>
<td>4</td>
</tr>
<tr>
<td>General operations</td>
<td>3</td>
</tr>
<tr>
<td>Education and training</td>
<td>3</td>
</tr>
<tr>
<td>Bolt-tightening</td>
<td>2</td>
</tr>
<tr>
<td>CNC machine operations</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 4 is used to highlight the overlap of industries and work processes within the reviewed papers. This chart highlights whether there is a more niche area that stands out in terms of active research, compared
to only looking at industry sectors and work processes separately. This also allowed us to review which areas overlapped across multiple industry sectors or work processes (see 9. Appendix, Table 35). It can be seen that “general assembly” at “sheltered workplaces” has the most repeated research. It can also be seen that “general production” has the most dispersed coverage, with research being discussed in general operations, general manufacturing, general assembly, education and training, and other/general work process categories.

Figure 4. Industry sectors vs. work processes.

2.3.3.2. Industry Sectors and Work Processes – Discussion

For the different industry sectors, a majority of papers were classified as “general production” because they provided higher-level frameworks, guidance for implementing GfM, or reviewed a broad body of literature. This category also includes empirical studies, but these empirical studies were used to produce frameworks and recommendations for the process of implementing GfM for a specific use case within production (e.g., MES systems). The sheltered workplace and automotive categories had the next highest count of papers, and this peak can be traced back to the focused nature of two separate groups of researchers. These researchers published numerous papers within the same topical area, thus they amassed a greater quantity compared to other categories which were one-off papers for several topics.

In the work processes categories, “general assembly” was the category with the most count of applicable papers. This is primarily attributable to the two major research projects included in this review – one set in the sheltered workplace industry, the other in the automotive industry. In both instances, the GfM scenario centered around assembly processes. Given this context, this cannot be seen as conclusive evidence of where GfM is most successful; however, it can be an indication that there might be areas where GfM is most easily testable and accessible for manufacturing applications. Also in the assembly
instances, the researchers were investigating assembly by humans, not robots or co-bots. Order picking was the next most classified category. Between assembly and order picking, there were 16 papers. These 16 papers account for nearly half of the papers within this research study (16 out of 35 = 46%). Without considering the human aspect of any other papers in this literature review, 46% of the papers included in this study focused on applying GfM in scenarios that required significant human work and interaction between human workers. Although much of production is moving towards Industry 4.0 ideologies with robots, co-bots, and various other automation technologies, there are many instances of production with significant human involvement that will require advancements for the human (i.e., the Operator 4.0 (Romero et al., 2016)). GfM provides an opportunity for advancements in processes that, for several reasons, are not or will not be fully or predominantly automated and will continue to rely heavily on human operators for their success.

Although gamification is commonly known for its applications in education and training (Keepers et al., 2020), this literature review only considered two papers that discussed education and training. This is primarily due to the scope of the paper and the strict definition and application of inclusion and exclusion criteria for the transparent classification of papers. The purpose of this literature review was to focus on papers related to shop-floor operations and tasks for manufacturing facilities.

Two papers were classified in the “general operations” category. To provide context for this category, these two papers are explained in a summary. One of the papers, (Elitok & Bolat, 2021), centers around applying gamification within the air transport industry and evaluates the impact of gamification based upon contributing factors of performance for operations within an airport. The other paper, (Schuldt & Friedemann, 2017), focuses on providing general recommendations for integrating GfM in various scenarios, thus it could not be classified as only education, general manufacturing, or general assembly.

In the review of the overlapping industry sectors and work processes (see Figure 4), the major findings are logical and unsurprising. For all instances in the automotive sector, the work processes are either assembly or bolt-tightening. This is mostly due to the one specific project based in the automotive industry, but also because these are the key processes involved in what is considered to be the automotive industry. The raw materials of the automotive industry which often involve other work processes, such as machining or rolling, are often considered to be part of the metal industry. For the sheltered workplace, all instances are for assembly. Almost all work processes in sheltered workplaces are assembly processes, and the papers based on research in the sheltered workplace were based around a specific research project. One other finding from the overlap chart for industry sectors and work processes is that the order picking was split between warehousing and general logistics sectors. This is logical and supports the fact that order picking is most frequently involved in these types of work environments.

2.3.4. Definitions of Gamification

The definition of gamification often resides in a gray area, with researchers using different definitions as a basis for their work. In many cases, researchers utilized an adapted definition of gamification – which is an indication of the early stage of research in which gamification finds itself, and how gamification differs from serious games, toys, and playful design. In an attempt to understand the similarities and differences as well as to identify the most generally accepted definition or at least core elements of currently in use definitions for gamification, Table 4 reviews the various definitions provided by the authors of each of the analyzed papers.
2.3.4.1. Definitions of Gamification – Results

Notice that the quantity of definitions does not add up to the total amount of papers reviewed (n=35). In six papers, their authors referenced more than one source when providing a definition. This is reflected in the counts in Table 4. Table 4 did not account for references that provided a chain-linked definition. For example, if an author only referenced paper X, and paper X then referenced the author Deterding, this citation was not added to the author Deterding, but as only a reference for paper X. If a definition was adapted from a reference and not used word for word, it was still cited as using that reference. “No definition provided” indicates that a paper did not explicitly state a clear and concise explanation for gamification. In most cases, they likely described gamification through a series of sentences or observations.

From these results, it is evident that the author Deterding has been established as the most widely accepted definition for gamification to date among the selected papers included in this literature review. This result is further supported by the many citations gathered by Deterding’s paper in which he defines gamification (Deterding et al., 2011), with over 9,505 citations today on Google Scholar (as of March 28th, 2022). Other references do not come close to the number of citations that Deterding received in the realm of definitions.

Table 4. Tabulation of key definitions in gamification research.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Qty</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Deterding et al., 2011)</td>
<td>21</td>
<td>Use of game design elements in non-game contexts.</td>
</tr>
<tr>
<td>(Kapp, 2012)</td>
<td>2</td>
<td>Using game-based mechanics, aesthetics and game thinking to engage people, motivate action, promote learning, and solve problems.</td>
</tr>
<tr>
<td>(Harris &amp; O’Gorman, 2014)</td>
<td>1</td>
<td>The presence or addition of game-like characteristics in anything that has not been traditionally considered a game.</td>
</tr>
<tr>
<td>(Huotari &amp; Hamari, 2012)</td>
<td>1</td>
<td>A process of enhancing a service with affordances for gamefication experiences in order to support users’ overall value creation.</td>
</tr>
<tr>
<td>(Kumar, 2013)</td>
<td>1</td>
<td>Application of game design principles and mechanics to non-game environments.</td>
</tr>
<tr>
<td>(Werbach &amp; Hunter, 2012)</td>
<td>1</td>
<td>Use of game elements and game-design techniques in non-game contexts.</td>
</tr>
<tr>
<td>(Chou, 2019)</td>
<td>1</td>
<td>Craft of deriving fun and engaging elements found typically in games and thoughtfully applying them to real-world or productive activities.</td>
</tr>
<tr>
<td>No definition provided</td>
<td>9</td>
<td>No definition was provided.</td>
</tr>
</tbody>
</table>

2.3.4.2. Definitions of Gamification – Discussion

One explanation for the overwhelming use of Deterding’s definition is likely its relative vagueness and thus broad applicability. Deterding’s definition is by far the shortest and does not provide any specificities of how to use gamification. All other definitions seem to build on Deterding’s and tie in additional criteria for defining gamification. Furthermore, Deterding’s definition was established in 2011, as one of the fundamental works on gamification. Meanwhile, the other definitions emerged later; in the order of Table 4, they were established in 2012, 2014, 2012, 2013, 2012, and 2019. Although many of the definitions were published closely to Deterding’s paper (2012-2014), it is understandable that these would have
fewer references because they are longer works, whereas Deterding’s paper is shorter, to the point, and easier to read – and states “defining ‘gamification’” explicitly in the title.

Based on these findings of the use of various definitions for “gamification,” it was concluded that Deterding’s definition emerged as an all-purpose definition that is widely accepted. Hence, the use of Deterding’s definition is recommended across the board to align the understanding and baseline of GfM research. The results show that this definition is well accepted by the community and is a testament to the broad nature and applicability of gamification. By narrowing this definition any further for a specific field, the creative power associated with gamification may be limited. While the alternative definitions referenced within this literature review provide truthful and applicable views for defining gamification, it is advisable to utilize a broad definition. Gamification provides a plethora of opportunities – some that may not have even been unlocked yet in the research world. Thus, it is important to not limit opportunity by using only a narrow definition at this time.

Of the seven definitions for gamification, four are from published books (Chou, 2019; Harris & O’Gorman, 2014; Kapp, 2012; Werbach & Hunter, 2012) and the remaining three definitions are from individual papers (Deterding et al., 2011; Huotari & Hamari, 2012; Kumar, 2013). This is significant in understanding the development of gamification from outside of the manufacturing research domain, where the majority of research is disseminated through research publications such as conference papers or journal articles. The heavy presence of published books highlights that gamification research today stems from, or at least has a strong footing in, other areas of interest, such as education, marketing, and others.

2.3.5. Game Design Elements
As Deterding suggests through his definition, game design elements are integral to implementing gamification. So, what are game design elements? These are the pieces of the scenario that create the “game-like” atmosphere.

2.3.5.1. Game Design Elements – Results
Of the 35 papers examined in this literature review, there were only four papers that did not explicitly state the specific game design elements they used. For the remaining 29 papers, all papers mentioned the use of multiple game design elements. It is noteworthy that no single paper mentioned only one game design element. In total, 64 unique game design elements were collected across the literature review (see Figure 5).
By focusing on game elements which were referred to at least ten times across the papers, six-game design elements are highlighted. In order of most common to least common, are (i) leaderboards, (ii) point/points/point systems, (iii) feedback (loops), (iv) badges, (v) levels, and (vi) progress bars. There was a total of 67 different game design elements identified. Of these, 32 were referred to only one time.

2.3.5.2. Game Design Elements – Discussions

It is interesting to note that no single paper focused on only one game design element exclusively. The inability to break apart individual game design elements for research is a known issue for gamification in manufacturing research (Bräuer & Mazarakis, 2019). This is a possible explanation for the lack of understanding of which game design elements are “best” or most applicable in various scenarios, as there are no controlled experiments where these game design elements have been evaluated individually and compared on their own. However, this is not unexpected given that gamification is a complex scenario that most likely benefits from one or more game design elements being applied simultaneously. Future research in GfM suggests the controlled experimentation of using different game design elements to identify the optimal mix of elements for particular work scenarios. This can be quantitative experiments but also qualitative research in the form of case studies to expand the current understanding of the interdependencies among different game elements.

The top six game design elements (see Table 5) derived from the literature are of no surprise to the research team. These elements are comparably easy to understand, easy to measure, and in some cases, easy to implement, especially compared to some other, more abstract or complex elements (e.g., stories, emotions, and loss aversion). These common game design elements work well for both areas of the gamification application in manufacturing (management and users/operators). As is often the case in gameplay, there is a set goal that is made apparent by playing, and these goals change depending on the game. For instance, in some games, the goal is to “be the best,” “to compete,” or simply “to win,” which applies to being at the top of the leaderboard. Points are often considered a suitable mechanism for differentiating between players’ abilities and performances. In other games, the goal may be to “get the
furthest,” which relates to the completion of levels, obtaining badges, and utilizing progress bars. Feedback and feedback loops apply to nearly every scenario where individuals are compared to one another or compared to a (fixed or adaptable) standard. For instance, when a person can achieve up to ten points, but they only achieve two points, this gives them feedback that they achieved very little compared to the fixed standard they are measured against in this scenario. This works similarly based on leaderboard rankings, obtaining badges, and leveling up.

Table 5. Description and implications of the top six commonly referenced game design elements.

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
<th>Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaderboards</td>
<td>A visual scoreboard showing identifiers (personal name, team, name, number, character, etc.) and current scores of the leading competitors.</td>
<td>Facilitates healthy competition among shopfloor workers (or teams) that helps increase progress towards a specific production-related goal.</td>
</tr>
<tr>
<td>Points</td>
<td>A mark or unit of scoring.</td>
<td>Provides a quantified overview of workers’ progress and abilities.</td>
</tr>
<tr>
<td>Feedback (loops)</td>
<td>Information about reactions to a person’s performance of a task is used as a basis for improvement.</td>
<td>Encourages continuous improvement by shopfloor workers by providing insights into errors or inefficiencies.</td>
</tr>
<tr>
<td>Badges</td>
<td>A distinctive emblem worn (physically or virtually) as a mark of membership or achievement.</td>
<td>May be used to distinguish between workers who have and have not achieved a particular training level or another set of experience/knowledge.</td>
</tr>
<tr>
<td>Levels</td>
<td>Each of a series of stages of increasing difficulty through which a player may progress, completing one stage in order to reach the next.</td>
<td>Showcases workers who have successfully completed a set of objectives, classified by the successful completion of a particular level.</td>
</tr>
<tr>
<td>Progress bars</td>
<td>A graphical element is used to visualize the progression of an extended task, sometimes accompanied by a percentage.</td>
<td>Motivates workers to continue working toward the completion of a task by highlighting how much they have already completed.</td>
</tr>
</tbody>
</table>

Of the 64 unique game design elements discussed herein, only 32 (50%) of these were mentioned in more than one paper. This is a testament to the flexibility and ever-evolving nature of gamification. This is also a testament to the broad scope associated with defining “gamification.” It may be beneficial to the GfM community to identify, cluster, and structure the distinctive design elements to aid in the selection of appropriate game design elements in support of an optimal mix of elements for implementing gamified solutions.

It is also important to consider which game design elements, or a mix of game design elements, when applied, create a gamified scenario. For instance, if only “feedback loops” are implemented, this likely does not imply a gamified scenario, as many people receive feedback on their work daily but would not consider it a gamified experience. On the other hand, if a variety of game design elements are implemented in a coordinated measure, the scenario may become more of a “serious game” and less of a gamified experience. In essence, there is a gray area that differentiates a gamified scenario from a not gamified scenario. The presence of one or more game design elements is not an exclusive criterion to
determine gamification. The coordinated and strategic application of game design elements is necessary for a scenario to be appropriately classified as gamified.

2.3.6. Referring to “Game Elements”
Game elements, as discussed in 2.3.5. Game Design Elements, have been referred to across different papers with various names, yet describe the same or similar ideas. In this subsection, the usage of various terms across the published works for game elements was reviewed, as well as how these names differ in their detailed, real-world definition.

2.3.6.1. Referring to “Game Elements” – Results
While Deterding clearly states the term “game elements” in his definition, other papers refer to these game elements with different phrases and terms. In some cases, authors may use different words interchangeably.

Overall, “elements” emerges clearly as the most commonly used term across the reviewed publications (see Table 6). Mechanics, dynamics, and components are the runner-ups when it comes to the number of uses for specific terms, while techniques, aesthetics, affordances, and technology were used less often. Of the 35 papers, nine reference something resembling the idea of game elements; however, they did not specifically identify them with a specific term. Considering this fact, 24 papers did use at least one broad term to describe game elements, and it is shown that 22 papers used the term “elements.” Thus, nearly all of the papers which use a broadly applicable term, also use the term “elements.”

Table 6. Alternative terms to refer to "game elements," the frequency of each in the dataset, and the definition of each.

<table>
<thead>
<tr>
<th>Term</th>
<th>Qty.</th>
<th>Definition (from Oxford Languages)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elements</td>
<td>22</td>
<td>• A part or aspect of something abstract, especially one that is essential or characteristic.</td>
</tr>
<tr>
<td>Mechanics</td>
<td>10</td>
<td>• The machinery or working parts of something.</td>
</tr>
<tr>
<td>Dynamics</td>
<td>5</td>
<td>• The forces or properties which stimulate growth, development, or change within a system or process.</td>
</tr>
<tr>
<td>Components</td>
<td>4</td>
<td>• A part or element of a larger whole, especially a part of a machine or vehicle.</td>
</tr>
<tr>
<td>Techniques</td>
<td>2</td>
<td>• A way of conducting a particular task, especially the execution or performance of an artistic work or a scientific procedure.</td>
</tr>
<tr>
<td>Aesthetics</td>
<td>1</td>
<td>• A set of principles concerned with the nature and appreciation of beauty, especially in art.</td>
</tr>
<tr>
<td>Affordances</td>
<td>1</td>
<td>• The quality or property of an object that defines its possible uses or makes clear how it can or should be used.</td>
</tr>
<tr>
<td>Technology</td>
<td>1</td>
<td>• The application of scientific knowledge for practical purposes, especially in the industry.</td>
</tr>
<tr>
<td>Not named</td>
<td>9</td>
<td>• n/a</td>
</tr>
<tr>
<td>Multiple names</td>
<td>9</td>
<td>• n/a</td>
</tr>
</tbody>
</table>
Additionally, of the 35 papers considered in this review, nine used multiple terms interchangeably. This is consistent with the fact that there is no concrete recommendation for how to refer to a broad selection of “elements.”

2.3.6.2. Referring to “Game Elements” – Discussion
Deterding has been successful in establishing a well-known and well-accepted definition for gamification. This definition has supported the most accepted use of the term “game elements.” However, in comparison to the acceptance of the definition, the term used to describe game elements is not as dominant over the alternate terms identified in the papers. The reason for this plethora of unclearly defined terms describing the game design elements is the perceived meaning of the specific words, as well as the lack of research deriving and defining the best term to be used. While Deterding clearly laid out the definition of gamification, similar actions have not been undertaken for defining game design elements.

A few select papers refer to a particular hierarchy for different terms. For instance, (Liu et al., 2018) refer to the game element hierarchy by Werbach and Hunter (Werbach & Hunter, 2012). This pyramid breaks the game elements into dynamics, mechanics, and components (see Figure 6). Descriptions for each level of the hierarchy are provided, but these do not align with the general term definitions described in Table 6.

![Figure 6. Pyramid of dynamics, mechanics, and components based on work from (Werbach & Hunter, 2012) as seen in (Liu et al., 2018).](image)

While there is no pressing need for the use of a single term here, there is a need to clearly define what each term means and encompasses in the view of GfM to ensure future (interdisciplinary) research can build on a solid foundation. This clarification should rely on generally accepted definitions of the terms. This aspect is vital to the growth of GfM so that individuals can understand the distinct aspects of gamification when they are working to implement it in a variety of different settings and applications. This is ever more important given the previously stated interdisciplinary nature of gamification research that includes engineering, computer science, business, and social sciences, just to name a few.

2.3.7. Technology Pairings
In the rapidly developing Smart Manufacturing and Industry 4.0 landscape, it is common to find an increasing number of different technologies in the workplace. Although gamification can be achieved
manually with a paper-and-pencil approach, GfM is more frequently implemented alongside technology. In the following, it was analyzed what technologies were used and associated with the GfM approaches in the identified use cases.

2.3.7.1. Technology Pairings – Results
Papers were categorized based on the type of technology that was used when implementing a GfM scenario (see Table 7). The non-specific category has the highest number of papers because it includes papers that focus on gamification in general use, without focusing on one or a few specific technologies. The most common technology for GfM is “computers,” which includes the use of computers for displaying information and integrations of gamification into Manufacturing Execution Systems (MES). Augmented Reality (AR) was commonly found across GfM applications, often in warehousing implementations for order picking scenarios. Motion recognition was commonly used as well to track progress towards a goal. This was important so that the gamification system could work autonomously for displaying feedback to operators. Virtual Reality (VR) was the focus of two papers. Projection, while only listed with a value of one, was seen across four papers in total. The difference between these two values is because three of the papers utilizing projection also utilized motion recognition and were categorized as such. This was due to the more technical nature of utilizing “motion recognition” compared to the simple display technology needed for “projection.” Smartphones and facial recognition technologies were included in two papers.

Table 7. Technology pairings.

<table>
<thead>
<tr>
<th>Technology Focus</th>
<th>No. of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer (incl. MES)</td>
<td>6</td>
</tr>
<tr>
<td>Augmented reality</td>
<td>5</td>
</tr>
<tr>
<td>Motion recognition</td>
<td>5</td>
</tr>
<tr>
<td>Virtual reality</td>
<td>2</td>
</tr>
<tr>
<td>Projection</td>
<td>1</td>
</tr>
<tr>
<td>Smartphone</td>
<td>1</td>
</tr>
<tr>
<td>Facial recognition</td>
<td>1</td>
</tr>
<tr>
<td>Non-specific</td>
<td>14</td>
</tr>
</tbody>
</table>

2.3.7.2. Technology Pairings – Discussions
Smartphones are possessed by nearly everyone in today’s society and are essentially powerful computers in pockets, equipped with cameras, large displays, and microphones/audio features. Yet, these are often frowned upon to be used in work settings, especially on a manufacturing shop floor, often for valid reasons such as safety considerations. It is interesting that considering the capabilities of today’s smartphones, only one GfM scenario reviewed for this literature review investigated the use of smartphones as an enabling technology for gamifying a manufacturing process. This is an area that is worth further exploration. A smartphone or tablet can be considered a more flexible and lower-cost investment than a computer system, easier to carry around, natively equipped with a high-resolution camera, microphone, and other features, and while still maintaining similar capabilities to a computer.

For the four scenarios that used projection, these are all papers written by Korn et al. based on a particular use case. While the projection applications seemed to be efficient with mostly positive results, this was only one use case. The use of projection needs to be further assessed in additional unrelated use cases and scenarios, particularly instances in manufacturing where the user is not near a clear and flat surface.
where projection would lend itself well to generalize the findings. Additionally, the cost and implementation of projectors are likely beyond that of computer or smartphone systems.

AR was one of the most used smart technologies across the reviewed papers. While this may be a likely beneficial use for gamification, it is important to note that majority of the papers within this category were implementations in a warehouse for the specific operation of order picking. Before producing guidelines for how to implement gamification using AR, it is important to explore the use of AR in other areas of operations, specifically production and manufacturing. In these facilities, AR may not be useful for several reasons, such as (i) obstruction by the AR glasses, especially when wearing safety glasses, (ii) expensive investment for the potential return, and (iii) often less looking and searching for objects involved in a repetitive manufacturing process compared to order picking.

In ensuring an autonomous, real-time gamification system, the system must be able to detect the actions of the worker, specifically for calculating metrics and measuring product quality. This is a great opportunity for the introduction of motion recognition to the manufacturing shop floor. The high rate of papers categorized in this literature review within the motion recognition category is understandable and exciting to see. One of the hardest aspects of ensuring a comprehensive and successful GfM system relies on reliable, accurate, and fast feedback to users, which in turn requires equal standards for the detection of progress. Looking at the results, motion recognition seems to be the best option for this necessary detection and tracking at this time.

In a general overview of how gamification was paired with various technologies within the papers reviewed in this study, there does not appear to be any primarily dominating technology required for implementing GfM. Given the current body of research, it is difficult to make recommendations on which pieces of technology to implement or use for a given GfM scenario. However, with additional research, it may show that AR is a feasible application for gamification in warehousing, while motion recognition is almost a necessity for gamification in hands-on manufacturing and assembly processes. Most importantly, the findings herein highlight the versatility of gamification to be applied with several types of technology (see Table 8).
Table 8. Technologies presently used alongside GfM and their descriptions and potential applications.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Description</th>
<th>Potential Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer (incl. MES)</td>
<td>The use of a computer system or monitor to display gamification, as well as the use of MES systems to provide data to the gamification scenario.</td>
<td>An MES module provided manual assembly data in near real-time and informs a human-centered performance management system which provisions information to the shopfloor workers (Hellebrandt et al., 2020).</td>
</tr>
<tr>
<td>Augmented Reality (AR)</td>
<td>A technology, usually glasses, which superimposes computer-generated images on a user’s view of the real world, providing a composite view.</td>
<td>AR glasses were used in a warehouse setting for order picking by providing pathways to specific shelves and highlighting the necessary objects on the shelves (Bräuer &amp; Mazarakis, 2019).</td>
</tr>
<tr>
<td>Motion Recognition</td>
<td>A technology that interprets human gestures to record the progress and actions of an assembly task.</td>
<td>Motion recognition was used to register workers’ movements and to then inform a gamified dashboard (Korn, Funk, et al., 2015).</td>
</tr>
<tr>
<td>Virtual Reality (VR)</td>
<td>A computer-generated, simulated experience, making the user feel they are immersed in the virtual environment.</td>
<td>VR was used to simulate a vehicle assembly process for training purposes (Butean et al., 2019).</td>
</tr>
<tr>
<td>Projection</td>
<td>The use of a projector to display an image on a surface.</td>
<td>A flat surface at the workstation allowed the gamified dashboard to be projected in front of users (Korn et al., 2017).</td>
</tr>
<tr>
<td>Smartphone</td>
<td>A mobile phone that performs many functions of a computer, typically with a touchscreen interface, internet access, and the ability to run downloaded applications.</td>
<td>An app for a smartphone was used to track and monitor CNC machine operators’ work processes and provided a gamified job design interface to the users (Liu et al., 2018).</td>
</tr>
<tr>
<td>Facial Recognition</td>
<td>A technology that registers and processes human faces using biometric data.</td>
<td>Facial recognition was used to register the emotions of workers during an assembly task to assess their feelings towards a gamified scenario (Korn, Boffo, et al., 2015).</td>
</tr>
</tbody>
</table>

Technology emerges as a driver and enabler of GfM with a diverse set of smart manufacturing technologies already actively applied. Table 8 illustrates solely technologies that were applied and/or described in the use cases of the reviewed literature. Given the rapid development of smart manufacturing technologies and the relative infancy of the GfM domain, it is highly likely that there will be even more diverse and broad utilization of technologies to enable GfM in the near future. This can encompass fundamental technologies such as Artificial Intelligence or more applied technologies such as Intelligent Virtual Assistants (IVAs). IVAs are conversational agents that can be introduced in a working environment to interact directly with the workers using natural language. As such, they can support the workforce in completing their assigned tasks by providing them instructions on how to complete a task or helping them with the tools they have to use for a task (Silva-Coira et al., 2016). It seems like a logical next
step to integrate game elements to guide the interaction of IVAs and human operators to improve interactions in the future.

2.3.8. Benefits
As stands true with any change, the introduction of recent technology, or the implementation of a new framework such as gamification in manufacturing settings, the question is always “why” – why should this be implemented and why is this solution valuable in this case? Often these types of questions can be answered by referring to the benefits associated with the particular solution. In this subsection, the focus is on the benefits associated with GfM, according to the literature considered in this review.

2.3.8.1. Benefits – Results
The most prominent benefits for GfM that are provided by the literature include improvements in (i) motivation, (ii) productivity and efficiency, (iii) engagement, interest, and focus, as well as (iv) encouraging and improving learning. Other notable benefits include increased enjoyment, flow, and social relatedness. The “other” category included in Table 9 accounts for benefits that were mentioned only once, such as voluntary completion of work (Liu et al., 2018), finding hidden strengths of employees (Korn, 2012), and providing insights to administrators (Lithoxoidou et al., 2017).

Table 9. Benefits of GfM as described in the literature, ranked.

<table>
<thead>
<tr>
<th>Count</th>
<th>Benefit Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>Motivation</td>
</tr>
<tr>
<td>23</td>
<td>Productivity and efficiency</td>
</tr>
<tr>
<td>16</td>
<td>Engagement, interest, and focus</td>
</tr>
<tr>
<td>14</td>
<td>Learning (encourage and improve)</td>
</tr>
<tr>
<td>10</td>
<td>Enjoyment</td>
</tr>
<tr>
<td>8</td>
<td>Mental flow</td>
</tr>
<tr>
<td>8</td>
<td>Social relatedness</td>
</tr>
<tr>
<td>5</td>
<td>Decrease monotony</td>
</tr>
<tr>
<td>5</td>
<td>Informative</td>
</tr>
<tr>
<td>4</td>
<td>Behavior change</td>
</tr>
<tr>
<td>3</td>
<td>Commitment and loyalty</td>
</tr>
<tr>
<td>2</td>
<td>Satisfaction</td>
</tr>
<tr>
<td>10</td>
<td>Other</td>
</tr>
</tbody>
</table>

2.3.8.2. Benefits - Discussions
The benefits described in Table 9 can be further categorized into two larger segments – psychological and production-related benefits.

The breakdown of the benefits into two large categories, psychological and production (see Table 10), provides multiple insights. Firstly, it is clear that the main benefits of GfM relate directly to psychological benefits. It is well established that an improved mental state can improve productivity. While focusing on improving mental states and psychological benefits for workers does not directly correlate to production benefits, production benefits are likely to occur, especially in the long term. Secondly, the fact that psychological benefits are more numerous than production benefits supports the notion of limited research on gamification in manufacturing. In manufacturing, the return on investment is often one of the
most crucial factors for implementing a change in the facility. Considering that it is nearly impossible to put a dollar sign on the benefits of improving the mental health of workers, it is more difficult for teams to explain the return on investment to implement gamification in manufacturing. Thus, there are limited opportunities for research on gamification in manufacturing. Determining how to quantify the benefits of GfM will be vital for explaining the cost-benefit ratios to manufacturers.

Table 10. Benefits of GfM in two large categories: psychological and production.

<table>
<thead>
<tr>
<th>Count</th>
<th>Category</th>
<th>Includes</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>Psychological</td>
<td>Motivation, engagement, enjoyment, mental flow, social relatedness, decreased monotony, commitment, satisfaction, and other (psychological).</td>
</tr>
<tr>
<td>51</td>
<td>Production</td>
<td>Productivity, learning, informative, behavior change, and other (production).</td>
</tr>
</tbody>
</table>

Recalling 2.3.2. Current Stage of GfM Research, which discusses the current state of gamification research for manufacturing and its lack of experimentation on the shop floor, one possible explanation for this lack of research in this domain may be due to the difficulty of measuring and experimenting with psychological benefits. In the manufacturing context, experiments are often set up and conducted differently from ones involving psychological components and human subjects. Many manufacturing and shop floor experiments focus on measuring and gathering primary (sensor) data and analyzing that data for production benefits (e.g., quality, efficiency, etc.). However, on the other hand, measuring and analyzing data for psychological benefits in manufacturing is often difficult, likely to involve more qualitative components, and the benefits of measuring psychological benefits are often indirect. Furthermore, the threshold of Institutional Review Boards (IRBs) is higher for experiments involving human subjects concerning personal data and allowable measurements.

Understanding and validating the urgent need and potential benefits that can come from improving workers’ mental health is necessary for the realm of production with the readily available capabilities of smart technologies. The reason for continuing to utilize human workers in production settings where robots have not yet, and will not in the near future, replace human workers is because of unique advantages: human ingenuity and problem-solving skills, among others. To ensure clear and productive thinking processes for human workers, their mental health is of utmost importance. In the future of production, the physical work will be primarily conducted by machinery with limited human involvement, while many cognitive functions will remain to be primarily conducted or controlled by humans. Therefore, production research needs to focus on methods that improve and maintain humans’ healthy mental state and thus ability to think clearly.

2.3.9. Limitations and Challenges

No solution approach is a one size fits all solution. When considering implementing something new to a manufacturing facility, it is vital to understand what may cause trouble and barriers. Additionally, there are always areas or problems that a single solution approach may not be able to solve. Given the results discussed in 2.3.8. Benefits, GfM faces not just technical challenges but also possible resistance from the operators themselves. Understanding the limitations and challenges of using GfM is crucial for anyone interested in gamification, practitioners and academics alike.
2.3.9.1. Limitations and Challenges – Results

Applicability and lack of research are the two most prominent limitations for GfM reported in the literature (see Table 11). Additionally, the corporate culture and ethics of GfM often impact the ability for gamification to be implemented and to work as envisioned. The remaining limitations, while small in the number of mentions within this set of papers, still provide significant barriers to those considering implementing gamification on their shop floor or beyond.

Table 11. Limitations and challenges of using GfM, according to the literature.

<table>
<thead>
<tr>
<th>Count</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Applicability</td>
</tr>
<tr>
<td>11</td>
<td>Lack of research</td>
</tr>
<tr>
<td>7</td>
<td>Corporate culture and ethics</td>
</tr>
<tr>
<td>5</td>
<td>Triviality</td>
</tr>
<tr>
<td>4</td>
<td>Recommendations</td>
</tr>
<tr>
<td>3</td>
<td>Requires updates and changes</td>
</tr>
<tr>
<td>2</td>
<td>Often requires more technology</td>
</tr>
<tr>
<td>16</td>
<td>No mention of limitations</td>
</tr>
</tbody>
</table>

The applicability category refers to the twelve papers stating that gamification is not applicable for a given scenario (six papers), process problem (four papers), or demographic of people (two papers). Lack of research refers to limitations which refer to limited research and guidance causing difficulty in knowing how to correctly implement gamification. The category corporate culture and ethics refers to overcoming the current status quo of the company as well as maintaining an ethical implementation of gamification and avoiding “exploitationware.” Triviality refers to gamification being a potentially less serious approach to solving a problem or being seen as only a short-term fix. Some articles provided recommendations instead of stating limitations by utilizing statements beginning with “gamification should...”. While this type of sentence is stated as a recommendation, it shines a light on the areas in which gamification is lacking and “should” be further investigated or challenged. A separate cluster was identified where the limitations specifically stated the need for reoccurring updates and changes for the gamified scenario to continue to have productive results. A couple of papers mentioned the need to implement more technology to use a particular gamification approach. Most notably, 16 papers did not discuss any limitations of GfM whatsoever.

2.3.9.2. Limitations and Challenges – Discussions

Consider how many papers do not explicitly state limitations or challenges associated with applying GfM. This may be seen as another indication of the explorative stage of research at this point in time. This is a key area to further the research and discussion of GfM across platforms. If limitations and challenges are not addressed appropriately, research has the potential to be unnecessarily duplicated. Also, this allows for research or implementations to occur in areas where GfM does not work well. Then, when it fails, it is given a “bad name” and often will not be tempted to be implemented again. If it is explicitly stated to not use gamification because of a specific limitation, that is better than no warning at all. It is imperative in manufacturing that appropriate tools are used for appropriate jobs. Gamification may be viewed as a tool for enhancing various aspects of manufacturing according to the benefits previously discussed in 2.3.8. Benefits.
While the literature has often emphasized the positive aspects and benefits of gamification, little emphasis has been put on the ethical use of gamification. This stands true for gamification in general, and even more so for GfM where ethical implications were regularly mentioned as a limitation (see Table 11), yet no single paper of the reviewed papers focused on the ethical implications in depth. For general gamification, according to (Shahri et al., 2014), “there is a fine line between being a positive tool to motivate employees and being a source of tension and pressure which could the affect the social and mental well-being within the workplace.” Hence, to avoid “exploitationware,” the GfM community must further explore and align the gamification strategies with the corporate culture, personality traits of the workforce, and managerial styles of leadership (Shahri et al., 2014). We further discuss our view on the ethical implications from a GfM lens in the conclusion section.

2.3.10. Future Research Recommendations
It is common to find suggestions for future work, based on the completed efforts, beyond what has been completed thus far. These future research recommendations often build off of a single paper’s recommendation. Here future research recommendations from the literature are reviewed and compiled into a more cohesive, unified list for the research domain of gamification in manufacturing. This provides the GfM field with a broad look into the future and ensures that the field is progressing in a forward motion. This also aims to ensure that researchers are not duplicating research.

2.3.10.1. Future Research Recommendations – Results
The future recommendations provided in this subsection are based on recommendations to further the field of GfM. In some instances, researchers provided future work recommendations that were specific to their research work only. For example, if a paper covers a single case study, the researcher may recommend that the future work focuses on duplicating the study or further analyzing the case study. While this may provide findings to the field of GfM, this does not directly supply insights into what is occurring with future work in the field.

The most common recommendation is to expand on different types of research (see Table 12). Of the 23 mentions for more research, 14 stated a need for more empirical research, six for more case study research, and three for more framework development. For the statements regarding empirical research, this was also broken down to specific needs within empirical research: seven for more time-based and longer timeframes of research, three for more people and larger samples, and three for more places such as different geographical locations or facilities. The next most-commonly recommended area of future research was to investigate the effects of GfM such as qualitative effects (six papers), quantitative effects (four papers), and more specifically novelty effects (two papers). The design outputs, such as different approaches or the use of different rewards, were mentioned in five different instances. The least commonly mentioned aspect to consider for future research are design inputs such as how the user affects gamification, specifically gender and demographics. Of the 35 papers, nine papers did not specifically mention any future research recommendations.
Table 12. Future research recommendations to further the field of GfM, according to the literature.

<table>
<thead>
<tr>
<th>Count</th>
<th>Future Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Type of research</td>
</tr>
<tr>
<td>12</td>
<td>Effects</td>
</tr>
<tr>
<td>5</td>
<td>Design outputs</td>
</tr>
<tr>
<td>2</td>
<td>Design inputs</td>
</tr>
</tbody>
</table>

2.3.10.2. Future Research Recommendations – Discussions

These future research recommendations align closely with the limitations described in the previous subsection. With a limited past, future research is an important field to ensure the growth and continued exploration of future research. It is evident that there is a need for additional empirical research given the current body of published research. This makes sense because that type of research often comes after the initial thought process of developing a new idea. GfM is currently at the phase of transitioning from the initial idea stage to the empirical and data-driven phase of further understanding gamification.

While there are significant findings from the future research recommendations from many papers, it is notable that nine of the 35 papers (26%) did not include any future research recommendations for the field of GfM. The growth of the field is dependent on expanding currently limited research. Providing sufficient recommendations to facilitate the future of the field is vital for sustained growth and a strong foundation to build on.

2.4. Conclusions

Gamification research is steadily growing and has begun finding its way to the shop floor and manufacturing systems. To capture the state-of-the-art of this emerging area in this paper, the results of our literature review centered around gamification applications for production, intralogistics, operations, and manufacturing within the walls of a factory. The collection and refinement of papers resulted in a total of 35 papers reviewed in-depth for the study. The analysis of the 35 selected papers lead to the following key findings:

- **Growth of the field** – In recent years, there has been unsteady and non-linear growth for GfM. However, it is evident that the field is trending positively with the highest count of papers in recent years 2019, 2020, and 2021.
- **Type of research** – Most GfM research is focused on experiments off the shop floor and conceptual design work. This is promising in building a strong base for future research and implementations of GfM.
- **Industry sectors and work processes** – GfM has shown research applications across a myriad of industry sectors and work processes, with no clear concentration in any one particular area. A key contributor to a focus in a particular industry or work process is dependent upon a research team’s chosen focus.
- **Defining gamification** – The definition formulated by Deterding is the most commonly used and accepted definition due to its broadness, allowing for interpretations to apply to various research areas.
• **Game design elements** – Across the literature for this review, 65 unique game design elements were identified, with half of these being referenced only one time, exemplifying the flexibility and broad applicability of gamification.

• **Referring to “game elements”** – The term “game elements” is the most commonly used term to describe the various ways to apply GfM; however, there are multiple other terms used across the literature.

• **Technology pairings** – Most commonly, GfM is implemented alongside a computer system. Smartphones are surprisingly used significantly less. Motion recognition and augmented reality are also commonly seen paired with GfM implementations.

• **Benefits of gamification for manufacturing** – Benefits can be broken into psychological- or production-oriented categories, with psychological benefits being the most prominently discussed benefit type.

• **Limitations of gamification for manufacturing** – The most common limitation of GfM is finding appropriate applicability and knowing when to use gamification for a given scenario or use case. Ethical considerations are common for general gamification, but, as of today, ethical concerns in GfM are discussed less frequently.

• **Future research of gamification for manufacturing** – Future research recommendations for GfM focused on expanding into different types of research, most notably a need for more empirical research.

Based on these conclusions, **four recommendations** are provided for the field of GfM:

i. **The use and acceptance of the Deterding definition** – to ensure that all research centers around the same topic and shares a similar fundamental understanding. Although Deterding’s definition, “the application of game design elements in non-game contexts,” can be considered to be comparably broad and vague, it accurately depicts the field of gamification in its ability to apply in various scenarios.

ii. **A clearer definition for the alternative game element terms** – such as mechanics, dynamics, etc. Although the general term, “game elements,” is necessary and sufficient for referring to elements in a broad sense, there is a need for more specific categories of the elements as well. These more specific categories for the elements would be beneficial for enhancing the field of GfM by allowing for a better understanding of how to implement or utilize each element. Specifically, it would be expected to see growth towards further defining, categorizing, and researching each element. The general term “game elements” is sufficient in referring to the elements in a broad sense; however, the accurate use of specific alternate terms will encourage enhancement in the field of GfM.

iii. **Additional empirical future research** – centered around understanding the true benefits of GfM, specifically how different game design elements impact the users. Diverse benefits are commonly mentioned for GfM, but few have been statistically evaluated or proven. Additionally, in many scenarios, applicability can be considered a current limitation of GfM. Designers of gamified applications struggle to identify game design elements that align with the requirements of and provide benefits for their specific use case.

iv. **The development of step-by-step guidelines for implementing gamification for manufacturing.** – More specifically, a set of guidelines is recommended with a focus on the different elements to be employed for implementing GfM based on different use cases and the potential benefits of
specific elements, as well as guidelines centered around the appropriate and efficient implementation process for different applications. This would be based on the recommended empirical research, allowing for a transparent and well-informed process for implementing gamification in the manufacturing industry. Furthermore, it will ramp up adoption by enabling production managers and other stakeholders in the industry to rethink their processes and participate in this field. Over time, this will provide further empirical data from real case studies contributing to the progress of the research field.

At this time, the limited research and results do not conclusively support nor oppose the use of GfM and its value for the realm of manufacturing. Although only a handful of positive results for the implementation of GfM have been described in research to date, equally few negative results have been reported. With that being said, there are many additional benefits commonly described for GfM compared to the number of limitations or challenges discussed in the available research. Gamification in manufacturing systems is still a research field in its infancy and significant research is desperately needed to evolve towards broad adoption in the industry.

Please note: This line denotes the conclusion of the material that has been acquired from (Keepers et al., 2022). The following material is new to support the dissertation document.

2.5. Research Problem
The overall goal of this dissertation research is to explore and establish GfM as a valid industrial productivity enhancement technique. Specifically, my research has focused on addressing all four of the previous research gaps through a succession of research tasks. Firstly, I have and continue to support the use and acceptance of Deterding’s definition of gamification by always using and citing this definition in all of my work. Secondly, through my publication, The Classification of Game Elements for Manufacturing (Keepers, M., Nesbit, I., Wuest, T., 2022), I utilized literature and clustered game elements into eight categories which supports the differentiation and comparability of various game elements. This is further described in the forthcoming subsection, 3.1. Aspects of the Gamified Scenario. Thirdly, the most significant portion of my dissertation research is the development of a thorough and detailed methodology paired with the complete data collection and analysis of an empirical study with human research subjects to address the identified gap of available empirical research on the topic of GfM. This dissertation serves to describe the entirety of the empirical study from motivation to conclusion throughout all of its sections. Lastly, my research supports the development of step-by-step guidelines for implementing GfM in two ways: i) by providing a framework alongside clear instructions for utilizing the framework and ii) by showcasing the validity of the framework through an in-depth methodology and analysis of a GfM mock assembly study.

2.6. Dissertation Research Overview and Hypotheses
My empirical study focused on assessing the instantaneous and long-term effects of GfM for a manufacturing assembly task. To simulate a manufacturing assembly task, I used the assembly of a Lego Telehandler (Lego Technic #42133). I gathered data regarding a common industrial key performance indicator (KPI), cycle time, which is the participants’ time to completion from start to finish. In addition to the KPI data, the modifying effects of feelings and perceptions of GfM and participant personality were assessed using the Myers-Briggs Type Indicator and the NASA Task Load Index (TLX) assessment. Significantly more details regarding the methodology for the research are included in the 4. Methodology section.
The aim of this empirical research is to assess the following hypotheses:

- **Production-oriented effects:**
  
  H1: GfM has no effect on the **cycle time** the **first time** an assembly task is completed.
  
  H2: GfM has no effect on **cycle time** of an assembly task **over time** (5 days, 2-3 hours per day).

- **Psychological-oriented effects:**
  
  H3: GfM has no effect on the **perceptions/feelings** of users the **first time** an assembly task is completed.
  
  H4: GfM has no effect on the **perceptions/feelings** of users for an assembly task **over time**.

- **Other related analysis of interest:**
  
  H5: A user’s **personality** has no effect on their **perceptions/feelings**.

The production-oriented effects focus on the cycle time of the assembly task. While I am aware that there are various industrial KPIs that could be of interest, I decided to primarily focus on cycle time. Cycle time for my application will account for both quality and efficiency of the process. When assembling the Lego Telehandler, there is a specified set of parts provided, so the build cannot be completed unless assembled correctly. Thus, the cycle time will encompass the time required for correcting any quality issues.

The psychological-oriented effects will focus on the perceptions/feelings of the users, especially their stress, confusion, and their affinity to using gamification. These items have been chosen as the focus of the psychological-oriented effects as they are commonly referred to in the literature as some of the key items of interest with the application of GfM.

Based on personal experience, discussions with experts, and our literature review, we believe that it is possible that a user’s personality may have an effect on their perceptions/feelings. This interaction of a user’s personality with their perceptions/feelings may have an impact on the results and conclusions of the psychological-oriented effects for the use of GfM. Thus, this interaction will be analyzed in conjunction with the study to provide a well-rounded analysis.
3. Background

This section illustrates the background of the experimental design, with a focus on tools and methods used for the study. First, we describe how the different aspects of the gamification scenario were established, particularly the game elements and the Lego set. Second, we describe why we chose particular variables to be of interest for the data collection.

The first subsection, 3.1.1. Game Elements, differs significantly from the remaining subsections of 3.1. Aspects of the Gamified Scenario. 3.1.1. Game Elements contributes to the results of the dissertation by providing information on the classification framework which works to address the fourth research gap, the development of step-by-step guidelines for implementing gamification for manufacturing. This contribution serves as an input for selecting the game elements, as described in 4.1. Preparation and Development.

3.1. Aspects of the Gamified Scenario

To ensure a gamified scenario, we stuck closely to Deterding’s definition for gamification: the application of game-like elements to non-game contexts (Deterding et al., 2011). As such, we had two major considerations when developing the gamified scenario, i) the game-like elements and ii) the non-game context. In this subsection, we describe the game elements that were considered for inclusion through an extended conversation on my classification framework (Keepers, M., Nesbit, I., Wuest, T., 2022). Then, we describe the non-game context that was utilized in our empirical study to simulate a repetitive and monotonous assembly task.

3.1.1. Game Elements

Please note: significant portions of this subsection have been acquired and adapted from my published, peer-reviewed article, titled “The Classification of Game Elements for Manufacturing Systems” (Keepers, M., Nesbit, I., Wuest, T., 2022), to fit the needs of my dissertation.

This subsection focuses on classifying the game elements often involved in GfM scenarios. The purpose of the classification framework is to provide a concrete method of identifying game elements to be used in and matched to gamified scenarios in manufacturing. Game elements are the attributes or pieces of the gamified scenario which create the game-like environment. To identify the game elements commonly mentioned in GfM research, the game elements were extracted from the results of a separate in-depth literature review conducted by our team (Keepers et al., 2022). A total of 35 papers, which focused on the gamification of shopfloor operations for manufacturing, production, and intra-logistics, resulted in a subset of 44 unique game elements to be considered in the development of the classification framework described in this subsection.

Following this introduction, 3.1.1.1. Methodology describes the methodology used for gathering the game elements from literature, as well as the development of the classification framework. 3.1.1.2. Results and Discussions provides the details and explanation of the framework. 3.1.1.3. Validation of the Framework showcases two different GfM scenarios in literature and how the framework can be used to develop a gamified scenario in manufacturing, thus validating the framework. The final subsection, 3.1.1.4. Conclusions, Limitations, and Future Work, completes the paper by drawing final conclusions and providing future work recommendations.
3.1.1.1. Methodology

This paper stems from an in-depth literature review previously conducted by the research team (Keepers et al., 2022). In the literature review, a methodical investigation was conducted to review all papers related to our topic of interest: GfM. We began the literature review by searching the Scopus database with the search string: TITLE-ABS-KEY((gamif*) AND (operation* OR manufactur* OR production* OR logistic)). After systematically reviewing the search results and identifying relevant papers, a subset of 35 papers were reviewed in-depth for the literature review. This literature review is not discussed further in this paper because, as it relates to this paper, it was used for the sole purpose of identifying the 44 unique game elements related to GfM. The purpose of this paper was to analyze the 44 unique game elements to establish a classification framework.

After reviewing the list of game design elements identified through our previous literature review, different categories were identified that the game elements could be grouped into based on their characteristics. The characteristics were categorized by simple vs. complex, individual vs. group, and intrinsic vs. extrinsic. These categories were chosen for the classification framework because the first two categories are important considerations of the scenario prior to implementing gamification, while the third category is more of a choice that can depend on personal preference or history of what has seemed to work well in the past. In general, the use of these three categories provides sufficient consideration of the current state (pre-gamification) of the scenario, as well as how the gamification designer and team desire the scenario to function. Additionally, all three categories included game elements which were classified into both categories.

When grouping between simple and complex game elements, a distinction was made based on the intricacy of creating or implementing the game element. For example, leadboards, point systems, and badges are most frequently used in a manner that only requires a relatively simple setup, allowing for universal use in a given scenario. For example, a leaderboard may display employee (or team) identifiers based on daily productivity in a manufacturing scenario where productivity is already a collected data point and reviewed metric. On the other hand, feedback, storylines, and levels often require more effort in the form of resources, time, and expertise for development and implementation, and in some cases may require personalization towards the user. For example, storylines, to be effective, require significant thought and development to ensure a clear and worthwhile result. Additionally, to provide automated feedback to users, it is likely that additional sensors and technology would need to be implemented into the scenario for data collection and accurate tracking of the employee’s or team’s work. In this comparison of simple vs. complex game elements, some elements were identified as being associated with both categories, where the complexity was dependent on the context in which it was used. An example of this is badges, as they could be made in a one-size-fits-all manner, or they could be individualized to each employee, team, or task.

Individual and group categorizations were decided based on the typical setting in which the game element is applied. Individual game elements can be applicable to a worker without the support or collaboration of anyone else. Quests, progress bars, and achievements are most often an individualized experience. On the other hand, group game elements include leaderboards, contests, and competition. These elements require more than one individual’s participation to be implemented successfully and effectively.

Intrinsic motivation involves the completion of a task because of personal satisfaction, while extrinsic motivation comes from outside motivators such as prizes, money, or social status. To rank game elements
between intrinsic and extrinsic motivations, it was important to take into consideration the differences between how they would be applied in a manufacturing scenario. Some elements were used as a means of enjoyment for an employee, whereas others may be used as an incentive in the workplace. For example, *strategies, avatars,* and *narratives* are a few game elements that primarily use intrinsic motivation. This is clear since these game elements do not lend to any external motivators. Contrarily, extrinsic motivators are elements like *leaderboards, levels,* and *rules.* These are known to be categorized as such because, when completed or used successfully, they lead to rewards such as increased social status, or other monetary and physical rewards. Some individual game elements could be considered either intrinsic or extrinsic motivators, depending on the context in which they were applied; *collaboration, interaction, teammates,* and *discussion boards* are a few game elements that could be considered either intrinsic or extrinsic depending on the context and objectives of the different scenarios. This was dependent on whether the collaboration was a personally rewarding experience or if it led to competitiveness, or other motivating factors.

Grouping game elements based on ‘technology required’ or ‘technology optional’ was considered, but after further assimilation, it was found that nearly every game element could belong in both categories. While technology would likely be helpful in the application of most game elements, the majority of game elements were still capable of being developed and implemented without the use of technology. For example, while *leaderboards* are most easily updated and displayed using smart systems, leaderboards can also be implemented in a manual sense where data is reviewed at the end of each time period (e.g., end of each shift) and handwritten and displayed on a white board on the shopfloor.

With these three categories identified as the basis for our assimilation, a tabulation was created that separated these categories into eight distinct groups of game elements. This tabulation created a framework layout to facilitate the classification of elements. With the classification of the game elements into the framework layout, the resulting tabulation allows for manufacturers, researchers, and designers to follow the framework to identify and select relevant, use case specific game elements when interested in implementing a gamified scenario into their operations.

Finally, based on the groupings and visualizations, a qualitative review of the findings was conducted. Using the findings and context of the use of game elements in GfM (through literature), the game elements were classified. This classification framework is the main result of this research and provides the groundwork for future implementations of gamification in manufacturing.

3.1.1.2. Results and Discussions

First, the classification framework (Figure 7) was developed by determining the most logical order of the columns as when reading left to right.

The first column is simple vs. complex since this is often the first stipulation of the scenario: determining and understanding the available resources of the project for implementing gamification. This is a crucial step where the designer would consider if there were sufficient time, money, expertise, etc. to warrant a complex scenario, as well as to consider if a complex scenario is currently worth the resource investment.

The second column is individual vs. group because this then considers the current state of the workforce by scrutinizing how the shopfloor employees interact and work with one another. In some cases, group work may entice and encourage additional efforts, while in other cases, group work may lead to arguments and stress.
Lastly, intrinsic vs. extrinsic motivation is considered as a third column. This was determined to be the last question to ask because this is solely based on the designer’s preferences, what they believe would be best, and how they desire the scenario to look and function.

![Classification framework depicting the eight groups for game elements.](image)

Following the completion of the framework, which resulted in eight discrete categories, the game elements were classified into their respective group(s) (Table 13). The results show a near even split of simple and complex motivators amongst the overall 44 game elements considered, with 24/44 (54.5%) classified as simple and 28/44 (63.6%) classified as complex. (Note that the percentages exceed 100%. This is due to the repeat of game elements in both categories.) This is beneficial to the gamification designer as the near even split indicates that making one choice (simple or complex) does not immediately rule out the possible use of a wide variety and selection of elements. This also helps to quickly narrow down the list of possible game elements, reducing the decision complexity at later states of the framework.

In the next step, far more game elements were classified as individual than group. 32/44 (72.3%) were classified as individual game elements, while 15/44 (34.1%) were considered to be group game elements. It is interesting to note that nearly all of the elements that fell into the group category were classified as such because they include a competitive element and thus contribute to extrinsic motivation because of a possible prize or other positive outcome, such as social status.

There are far more extrinsic (33/44 = 75%) motivators amongst the game elements than intrinsic (16/44 = 36.7%). Implementing extrinsic motivators in GfM creates a simpler and more easily measurable study for developers and implementors, because people tend to respond much faster to rewards that are offered to them than they do to intrinsic motivators. Many of the game elements that were classified in both categories were dependent on the context in which they were applied and how the individuals may interpret the tasks. For example, having teammates could be either intrinsic or extrinsic, because an employee may simply find personal enjoyment from the social aspect of teammates, but the employee may also find outside motivation of social status when they are seen as the team leader.
3.1.1.3. Validation of the Framework

To consider the efficacy of the framework, two publications which involved detailed scenarios of GfM implementations were reviewed in parallel with the framework. Using the understanding of the scenario of the first publication (Ohlig et al., 2021), we worked through the framework to determine possible game elements to be used. With the second publication (Korn, Boffo, et al., 2015), we considered the chosen game elements used in the paper alongside the framework to show that the classification framework can be used for considering additional game elements that could be used in a similar scenario.

One publication (Ohlig et al., 2021) observed a lab-based environment in which two teams completed a three-step assembly process on an assembly line. The first team worked on the assembly line with gamified information, including a leaderboard, where each employee got a score for the number of products they assembled and allowed communication amongst employees, while the second team did not receive a score and were not allowed to communicate. The purpose of the game was to assemble as many test objects as possible within the allotted time frame, completely and accurately. The group that had a leaderboard to track the number of completed product produced significantly more than the group that did not. Comparing the work of this publication with the developed framework, it is clear that leaderboards were used in a simple and group environment. Based on the reactions of the gamified team, it is shown that extrinsic motivators of the leaderboard were well received due to the gamified team’s discussion of their status on the leaderboard during the break. By understanding the specific situation that the researchers had (simple, group, extrinsic), our framework would have led them to using leaderboards to incorporate a gamified experience into their production processes. This evaluation result indicates that our framework is aligned with recent implementations and experimental outcomes.

In another study conducted by Korn (Korn, Boffo, et al., 2015), emotion recognition is used to determine whether employees in a production work setting were “happy” or “unhappy” when applying gamification methods. They were given a 10-task repetition without gamification, then the process was repeated with gamification. Participants were given real-time feedback, received performance displays, and saw their

Table 13. Game elements classified into their respective group.

<table>
<thead>
<tr>
<th>Group</th>
<th>Descriptor</th>
<th>Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>SII</td>
<td>Notifications, personal goals, increasing complexity, constraints</td>
</tr>
<tr>
<td>Group 2</td>
<td>SIE</td>
<td>Progress bars, goals, strategy, time constraints, increasing complexity, loss aversion, displaying performance, badges, rewards, achievements, awards</td>
</tr>
<tr>
<td>Group 3</td>
<td>SGI</td>
<td>Collaboration, interaction, teammates</td>
</tr>
<tr>
<td>Group 4</td>
<td>SGE</td>
<td>Displaying performance, collaboration, interaction, leaderboards, competition, teammates, ranking, awards, point system, scoring system, contests</td>
</tr>
<tr>
<td>Group 5</td>
<td>CII</td>
<td>Constraints, story elements, avatars, emotion, narrative, problem solving, puzzle games, scenarios</td>
</tr>
<tr>
<td>Group 6</td>
<td>CIE</td>
<td>Badges, rewards, achievements, awards, levels, feedback, quests, virtual goods, boss fights, challenges, mission, rules, content unlocking, performance groups</td>
</tr>
<tr>
<td>Group 7</td>
<td>CGI</td>
<td>Discussion boards</td>
</tr>
<tr>
<td>Group 8</td>
<td>CGE</td>
<td>Awards, point system, scoring system, contests, performance graphs, discussion boards, betting, social recognition</td>
</tr>
</tbody>
</table>
scores. Using our proposed framework, it is shown that point systems are associated with group 4 (simple, group, extrinsic). Following this, it is likely that teammates, rankings, and leaderboards would have had the same effect on the process as the chosen game element. Feedback is a group 6 (complex, individual, extrinsic) element, which suggests that game elements such as levels, badges, and rewards may have had similar effects on the process. This aspect of the validation process expands on the possible use of the proposed framework as a tool for implementing additional game elements that are comparable to game elements already used in a scenario.

In addition to validating the framework with two publication case studies, this dissertation research works to validate the classification framework through an experimental use case. The classification framework was utilized for the experiment to identify which game elements were most applicable to the scenario. This is further detailed in the 4.1. Preparation and Development.

3.1.1.4. Conclusions, Limitations, and Future Work

The formation and organization of this chart creates a foundation for those in industry and research to have a basis of application for GfM by providing guidance on the game elements to use. Since gamification is a growing application in manufacturing, this framework facilitates the future growth by providing concrete guidelines for implementation by considering the current state (pre-gamification) of the process.

By compiling a list of 44 game elements used across various publications, it was determined that the game elements could be separated into groupings of simple or complex, individual or group, and intrinsic or extrinsic. Using these categories, eight groups of differing game elements were created. The classification framework was validated using two different publications of GfM scenarios to exemplify two ways the framework can be utilized by stakeholders: i) in examining the pre-gamification scenario to understand potentially relevant game elements for the specific use case, and ii) in understanding the current gamified scenario to consider additional game elements which may produce similar results.

One limitation of this research is that the publications used for the validation step were part of the 35 papers from which the 44 game elements were extracted from. Thus, it is possible that a slight bias was introduced since the game elements and framework application for these publications were not entirely independent of one another. However, since the game elements were only extracted and then compiled separately into the proposed framework, the potential bias is limited. Another limitation of the research at the time of publication is that the evaluation is based on secondary data (publications). This dissertation aims to validate the proposed framework using primary research data (case studies or experiments). Additionally, the framework included in this paper does not include considerations for making decisions based on characteristics of the production system or operators involved, but only considerations which are under the control of the gamification designer. These limitations are addressed in the planned future work.

The proposed framework allows for future manufacturers to apply gamification techniques into their production processes and for GfM researchers to further explore gamification by understanding the benefits of each group of elements for manufacturing. Following the creation of this framework, the next step in the process is to validate the framework with industry partners through a case study. From this, adaptations can be made to improve the organization and implementation of the process. Additionally, one area that GfM struggles is in understanding the concrete, empirically-based benefits of gamification. This framework hypothesizes that benefits for elements within the same group are comparable, and thus the empirical testing of GfM benefits could be centered around the eight groups, as opposed to the 44
3.1.2. Lego Set

The focus of this dissertation is on repetitive and monotonous manufacturing assembly tasks. To simulate this type of assembly, we decided to use a Lego assembly kit. Lego kits (and similar assembly kits) have been used in multiple research studies to simulate an industrial assembly process (Butean et al., 2019; Korn, Boffo, et al., 2015). For our scenario, we utilized a Lego kit from the Technic line, specifically the Lego Technic Telehandler (kit #42133). These build kits produce a final product that has some mechanical mobility (steering and raising/lowering the boom) and resemble a reasonably complex technical system with multiple components. Additionally, the Lego Telehandler required roughly 30–60 minutes on average for a first time build by an individual with average experience in building Lego systems. This fit well with our desired timeline for the study, especially for the time requirement for participant involvement. The 30–60-minute average build time was established through ad-hoc assembly with fellow researchers prior to the study design to determine a rough estimate of build time. This data was not included as part of the empirical study and was used solely for selecting a Lego assembly kit with an appropriate build time for the study. We were also conscious in choosing a Lego kit with reasonable complexity in order to mentally stimulate participants during the build. During the decision process, we determined our experimental design required a Lego kit with a longer cycle time (in the 30–60-minute range for first build attempts) in order to minimize human error in the start and stop of the timer, as well as to allow for more obvious identification of differences in means.

Figure 8. Packaging for Lego Technic kit #42133, Telehandler used in this research.
3.2. Variables of Interest for Data Collection

In this subsection, we describe the different variables that were of interest for the research study. More information on how the data was collected for each of these variables is provided in 4.5. Data Collection.

3.2.1. Cycle Time

Cycle time, which refers to the time to complete the assembly of a single unit, was used to track the productivity of participants. During the design phase, we determined that measuring quality or accuracy of the build would be nearly impossible. We thought of a few situations that may occur that would make having a consistent quality measurement difficult: i) participants making minor mistakes vs. making major mistakes, ii) participants realizing their mistakes quickly vs. realizing their mistakes many steps later, iii) different methods of fixing mistakes (disassembling completely, pinpointing the mistake and fixing it directly without impacting other steps, or a mixture of techniques), and iv) participants’ ability to locate an error. With so many potential immeasurable impacts to quality, we elected to only measure cycle time and that the researcher would watch each build carefully and require 100% accuracy to allow a build to be considered complete. By requiring 100% accuracy, the cycle time acts as a measure of both accuracy and timeliness in the build, thereby representing productivity and efficiency in one uniform metric with no researcher bias.

3.2.2. NASA Task Load Index

To measure the perceptions/feelings of participants, we elected to utilize the NASA Task Load Index (TLX) assessment tool. The assessment was developed by researchers under the NASA umbrella in the 1980s and has since “become the gold standard for measuring subjective workload” for various fields and applications\(^1\). The NASA TLX focuses on six dimensions of perceived workload: mental demand (MD), physical demand (PD), temporal demand (TD), performance (P), effort (E), and frustration (F). These six dimensions worked well for the purpose of this study by providing a well-rounded analysis of the perceived workload, both physical and mental, that participants experienced. To measure the perceived workload for each of the categories, participants were asked to respond to the below questions on a scale from 0 to 100, in increments of 5:

- **Mental demand (low/high):** How much mental and perceptual activity was required (for example, thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, forgiving or exacting?
- **Physical demand (low/high):** How much physical activity was required (for example, pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
- **Temporal demand (low/high):** How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
- **Performance (good/poor):** How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
- **Effort (low/high):** How hard did you have to work (mentally and physically) to accomplish your level of performance?

\(^1\) TLX @ NASA Ames - Home
• Frustration level (low/high): How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

3.2.3. Personality

To determine a participant’s personality, we employed two methods. The first method we used was the Myers Briggs Personality Type Indicator (MBTI)\(^2\). The MBTI is a well-known, widely utilized, and free personality assessment for the general public that returns a four-letter acronym representing an individual’s personality. The first letter is always either an E or an I, relating to an extrovert or introvert personality type. The second letter is always either an S or an N, indicating sensing or knowing personality type. The third letter is always either a T or an F, indicating a thinking or feeling personality type. The fourth letter is always either a P or a J, indicating a perceiving or judging personality type. The MBTI also receives criticism in research communities as it has become rather commercialized, leading some to believe that its results are not completely accurate. Even knowing these criticisms, we felt the MBTI would provide a sufficient way of identifying general characteristics of our participant’s personality, especially for gathering information about their introvert/extrovert qualities since this relates back to the classification of game elements framework (3.1.1. Game Elements).

The second method we used to determine attributes of a participant’s personality was three survey questions. Two of the survey questions asked participants how they would rate their competitiveness in two situations, one being against themselves and the other being against other individuals. The third survey question asked participants how patient, or tolerant with difficult tasks, they believed themselves to be. These three questions were formulated as part of the end-of-study exit survey based on observations by the researchers during the earlier portions of data collection.

\(^2\) The Myers & Briggs Foundation - MBTI® Basics (myersbriggs.org)
4. Methodology

The purpose of this section is to describe how the research was conducted and the study set up, as well as to provide a background concerning why certain decisions were made. The methodology covers the process from preparation and development of the gamified assembly task to the completion of data collection. The section is structured as follows. First, we discuss how the gamified scenario was developed, as well as preparation of the gamification interface. Then, a brief overview of the experimental setup is provided before we dive further into the steps taken during the study. Next, we discuss the recruitment process of obtaining participants for the study, and this then leads into how we selected our participants. Following, we describe the data collection process wherein participants attended sessions and completed the mock assembly task. Lastly, we describe the exit survey that we implemented to help us gather a few final questions that we had for participants.

4.1. Preparation and Development

Prior to beginning any development or involvement with participants, since this study involved human research participants, we had to submit our detailed research plan and methodology to the West Virginia University Institutional Review Board (IRB). Our research plan did not involve any identifiable data and all personal data was deidentified using a coding system. Hence, we submitted the study to the IRB under the FLEX/NHSR protocol. Ultimately, we received approval for our IRB protocol, Protocol #2209651754, and began developing our interface, described below, and recruiting participants, described in the following subsection, 4.3. Recruitment.

Utilizing the aforementioned classification framework (Figure 7), described in detail in 3.1.1. Game Elements, we determined the most applicable scenario for this study would be a simple (S), individual (I), and extrinsic (E), or SIE, scenario. This led us to choosing from game elements in Group 2. The options in Group 2 were: *progress bars, goals, strategy, time constraints, increasing complexity, loss aversion, displaying performance, badges, rewards, achievements, and awards* (Table 13). Consequently, we decided to utilize *progress bars and badges* for our study as these were some of the most commonly discussed elements in literature, applicable for our scenario, and simplest to implement technically. We utilized Microsoft (MS) PowerPoint to develop our gamification interface, alongside the Apple built-in clock as a timer on an iPad. In MS PowerPoint, we created a file that, when changing from one slide to the next, it updated the screen to show an additional green box with the now-current step number (see Figure 9 and Figure 10). The design was purposely simplistic to be not distracting and to allow the participant to focus. At the start and end of each build, earned and to-be-earned badges were displayed to users (see Figure 11 and Figure 12).
Figure 9. An example screen of the progress bar which would be shown while a participant is working on step 16.

Figure 10. An example screen of the progress bar after successful and accurate completion of step 16, indicating that the participant is now working on step 17.
The badges we implemented included a combination of finite and infinite badges. The finite badges refer to badges which can only be achieved a finite number of times; in our case, this was one time. Finite
badges were used to motivate participants to return to the study for subsequent builds, as well as to provide them with the opportunity to earn badges regardless of their improvements.

Infinite badges, badges that can be earned as many times as possible in a timeframe, were used to allow participants the ability to earn badges based solely on their efforts and improvements throughout the study. The infinite badges also motivated participants to improve their times iteratively throughout the study. Both badge types are common in game play and are commonly implemented concurrently to provide incentives to users in multiple ways.

We included the following finite badges: Build #1 Complete, Build #5 Complete, Build #10 Complete, and Build #15 Complete (Figure 13). These badges were awarded when a participant completed the Lego build for the n’th instance, where n is 1, 5, 10, or 15.

![Figure 13. Finite badges awarded to participants.](image)

In our study, we included one infinite badge. The infinite badge available to participants was the New Personal Record badge (Figure 14) which was earned any instance which the participant completed a build faster than they had across any of their previous attempts.

![Figure 14. Infinite badge awarded to participants.](image)

When a participant earned a finite badge, the gamification platform would have the following text underneath the badge, “Earned on Date.” When a new personal record was achieved, the gamification platform would include the earned-on date, how many times the new personal record had been earned, and the new personal record time.

Additionally, the badges would be shown to participants in gray scale until they earned the badges. Once earned, the badges would take on their colorful version. This highlighted to participants which badges were available to them as well as showcasing their accomplishments, ultimately motivating them to continue to progress and improve throughout the study.
4.2. Overview of the Experiment

The experiment was primarily broken into three phases, i) recruitment, ii) participant selection and random assignment, and iii) data collection (Figure 15), which spanned across five separate visits for each participant. These three phases are discussed in depth in the following subsections.

![Figure 15. Overview of the methodology for the experiment.](image)

4.3. Recruitment

To collect primary research data to answer our hypotheses (H1-5), we recruited suitable participants into the study. To facilitate recruitment, those interested in participating in the study were asked to complete a recruitment survey via Qualtrics. Paper flyers with a QR code that linked to the survey, a survey link, and tabs that could be removed by people to take a copy of the survey link with them (Figure 16) were posted around campus and an email was sent out to the entire student body to advertise the study (Figure 17). A total of 213 responses were collected from the recruitment survey.
Figure 16. Sign posted around campus to recruit participants.
Participants needed for: Study on Effects of Changing an Assembly Task

Students at WVU who are 18 years of age or older are invited to participate in a study where they can earn up to $50 for their participation. The study will be held in the Engineering Sciences Building and involve 5 visits, at approximately 1 hour each visit, over the course of ~5 weeks. Participants will complete a Myers Briggs personality assessment, complete 15 builds of a Lego assembly, and complete other questionnaires. Please complete this form if you are interested in participating: https://wvu.qualtrics.com/jfe/form/SV_cNHJoDwyLniyW

Questions, comments, or concerns can be directed to Makenzie Dolly (Keepers) at mk0004@mix.wvu.edu.

The recruitment survey consisted of the following questions:

- **For contacting participants:**
  - First name
  - Last name
  - Email
- **Demographic information:**
  - Age (in years)
  - Current role
  - How would you rate your experience/knowledge with building Legos?
  - (Optional) Tell us more about your experience/knowledge with building Legos.

4.4. Participant Selection

From those who responded to the survey and showed interest in participating, using a random number generator, we contacted a total of 45 participants, ultimately ending with 20 participants to complete the entire study. Our goal was to have a total of 20 participants complete the study, as we believed this sample size would be sufficient for drawing meaningful conclusions while also keeping a manageable workload and timeframe for data collection. Additionally, the sample size of 20 was found to be the median (see Table 14) across other recent studies of GfM, such as those included in (Keepers et al., 2022).
Table 14. Sample size of experiments as seen from articles considered in the State of the Art.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Industry/Topic/Focus</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Lee et al., 2016)</td>
<td>Automotive bolt-tightening</td>
<td>10</td>
</tr>
<tr>
<td>(Roh et al., 2016)</td>
<td>Automotive bolt-tightening</td>
<td>5</td>
</tr>
<tr>
<td>(Seo et al., 2021)</td>
<td>Automotive bolt-tightening</td>
<td>18</td>
</tr>
<tr>
<td>(Liu et al., 2018)</td>
<td>CNC machine operators</td>
<td>60</td>
</tr>
<tr>
<td>(Putz et al., 2019)</td>
<td>Order picking</td>
<td>17</td>
</tr>
<tr>
<td>(Korn, Boffo, et al., 2015)</td>
<td>Assembly</td>
<td>10</td>
</tr>
<tr>
<td>(Ohlig et al., 2021)</td>
<td>Assembly (vocational training)</td>
<td>45 groups (135 trainees)</td>
</tr>
<tr>
<td>(Lessel et al., 2016)</td>
<td>General manufacturing</td>
<td>20</td>
</tr>
<tr>
<td>(Bräuer &amp; Mazarakis, 2019)</td>
<td>Order picking</td>
<td>63</td>
</tr>
<tr>
<td>(Butean et al., 2019)</td>
<td>Assembly</td>
<td>2 groups of 10 (20 total)</td>
</tr>
<tr>
<td>(Korn, Funk, et al., 2015)</td>
<td>Assembly</td>
<td>24</td>
</tr>
<tr>
<td>(Korn et al., 2014)</td>
<td>Assembly</td>
<td>40</td>
</tr>
<tr>
<td>(Korn et al., 2016)</td>
<td>Assembly</td>
<td>5</td>
</tr>
<tr>
<td>(Ponis et al., 2020)</td>
<td>Order picking</td>
<td>29</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

We contacted and coordinated with a greater number of participants to account for those who may no longer be interested, unable to commit to the entire study, or whom had to drop out prior to completion of the study. The data included in the analysis of this study considers the 20 complete samples. A complete sample is one wherein the participant completed the recruitment survey, the MBTI assessment, all 15 builds of the study, the five NASA TLX assessments, as well as the exit survey.

4.5. Data Collection

For the data collection phase, as a brief overview, participants were randomly assigned to a treatment group, either control or gamified. Once in their treatment group, they were required to complete a MBTI assessment, 15 mock assembly tasks and five NASA TLX assessments distributed across five separate visits, and an exit survey. In the following paragraphs, this process is described in greater detail from start to finish.

Once participants were accepted into the study, they were randomly assigned to either the control group or gamified (treatment) group. This assignment was not explicitly disclosed to them, and they were not informed that there was an “other” group participating in the study. During a participant’s first visit, they first had to review the consent form. Then, they were asked to refrain from discussing the study until its completion. Next, they utilized an iPad to complete a MBTI assessment using the application titled “The Personality Types” (Figure 18). This app was chosen as it was free to use, had a high rating in the Apple store with 4.8/5, and provided numerous reputable references within the app to cite its development. After the participant completed the MBTI assessment, their 4-letter coded result was recorded in MS Excel alongside their participant identification (ID) number.
Next, participants were provided with a completely disassembled Lego Telehandler set (set #42133) and the accompanying Lego build instructions manual (Figure 19).
The researcher then read the following explanation to participants to ensure all participants had a sufficient knowledge of how to utilize the build instructions:

“To complete the Lego assembly, you will use the user manual that is provided with the Lego kit. You may already be familiar with this, but we will go over a few key concepts so that all participants have the same information. *Go to page 4 of handbook* At the top of each page there is a white box that contains the images of the pieces that you will need to complete that step. On this page, at the bottom left corner, you see a Lego piece and a 1:1 ratio symbol. This means that the piece you are looking for would be this exact size. Only some steps will have a 1:1 image of a piece. The box with the number 5 in it indicates that the piece you are looking for is 5 blocks in length. You will see these number indicators throughout the handbook, but not for every piece, to help ensure you use the correct piece. The colors in the diagrams are the colors of the pieces you will use. As soon as you complete a step, please move forward onto the next step. You should use all of the knowledge that I just shared to help you complete the build efficiently. If at any point, you realize you made a mistake, please work to resolve the mistake then continue building. The build is not complete until it is done so correctly. Do you have any questions?”

For gamification participants, the gamified platform was displayed starting with the “Your Badges” screen showcasing all badges available for participants to earn, and an iPad with a timer was provided below the “Your Badges” screen (Figure 20). Then the following explanation would be read aloud to participants:

“In front of you, you have a large screen and an iPad. The iPad will display a stopwatch that indicates how long you have been building. The screen will begin by showing the “Your Badges” page. Currently, because you have not completed any builds, you have not earned any badges. If you have any questions about what the badges mean, please feel free to ask. When you are ready to start building, the screen will change to show a progress bar. Throughout your build, you will see the colors green, yellow, and gray. Green indicates the step is complete, yellow indicates this is the step you are currently on, and gray indicates steps that still need to be completed. A number will show in the yellow box which tells you which step number you are on, corresponding to the Lego build manual. If you make an error in your build, the progress bar will not move to the next step until the step has been completed correctly. At the end of each build, you will be notified of any badges that you earned, as well as provided a screen to review all of your earned badges from throughout the study. You do not need to click or touch anything on the large screen. When you have completed each build, press the “Stop” button on the iPad timer. Before we get started, first take a moment to review the badges on the screen. Then, when you are ready to begin building, please press the “start” option on the iPad timer and begin building. Do you have any questions?”
Once the appropriate explanations to participants (Lego explanation for control group, and Lego and gamification explanation for the gamified group) was completed, participants were instructed to begin building the Lego Telehandler. At the start of each build, a timer was started to track the cycle time of each build for each participant. The researcher overseeing the build watched intently to ensure the accuracy of the build but did not say anything or provide any corrections to the participant. For gamification participants, the researcher advanced the progress bar when the participant completed a step correctly. If the step was not completed correctly, the progress bar would remain at the step where the issue occurred. In a few instances participants in the gamified group would point out the progress bar to the researcher and question why it was seemingly behind/delayed where they thought they were in the progress of the build. The researcher would respond by stating, “the progress bar is accurate and up to date” without providing additional context. The time clock was stopped when the participant completed the build correctly. If the participant believed they completed the build, but had not done so correctly, the researcher would state, “one or more errors remain.”

After a build was completed correctly, the resulting time was recorded. For all participants, the completed Lego set was removed from the participant’s workspace, the Lego build manual was returned to the first page, and a new disassembled Lego kit was laid out randomly in front of the participant. Then, for gamification participants only, the badges were updated and displayed. Again, the participant built a Lego Telehandler in the same fashion as previously described. Once the participant completed the next build, the time was recorded, and the completed build was removed and replaced with a third disassembled set (last build of the visit). The build processes for the control group and gamified treatment group are depicted in Figure 21 and Figure 22.
Figure 21. Build process for control participants.

- Researcher instructs participant to start building and starts timer
- Complete step according to Lego build manual
- Indicate build is complete
- If accurate, researcher stops timer
- If inaccurate, press start button on timer, complete step according to Lego build manual, review badges on screen, and check gamification interface to ensure accuracy.

Figure 22. Build process for gamification participants.

- Review badges on screen
- Press start button on timer
- Complete step according to Lego build manual
- Check gamification interface to ensure accuracy
- Press stop button on timer
- Review badges on screen
Then, after completing three builds, the participant was asked to complete the NASA TLX assessment using the NASA TLX app from the Apple app store (Figure 23). Responses for the NASA TLX assessment were recorded into the data collection MS Excel sheet.

For visits two through four, participants were asked to complete the three builds according to their randomly assigned group designation and complete the visit by filling out the NASA TLX assessment. For the fifth visit, participants again completed the three builds and the NASA TLX assessment. At the conclusion of their fifth and final visit, they completed an exit survey.

4.6. Exit Survey
The purpose of the exit survey was to gather additional information that the research team was interested in obtaining from participants. This was based on comments from participants throughout the study (visits
and other observations during experimentation. All questions for the exit survey, except for participant ID#, offered responses on a 5-point Likert scale. The following questions were included on the exit survey:

- Q1: Participant ID#
- Q2: How competitive would you consider yourself against yourself (you vs. you)?
- Q3: How competitive would you consider yourself against others (you vs. them)?
- Q4: How much pressure did you feel during the experiment to complete the build correctly?
- Q5: How much pressure did you feel during the experiment to complete the build faster each time?
- Q6: How patient are you with hard tasks?

Then, the exit survey had a screen instructing the users to notify the researcher when they reached the break in the survey. At the break in the survey, the researcher shared details with the participant regarding the “other” group of the study (the one they were not associated with). For control participants, a sample gamification interface was shared and explained. For gamification participants, the researcher explained that the other participants in the study had no interface in front of them. At no point in the study was the term “gamification” used. This was purposeful in order to deter participants from looking at the gamification scenario as more fun or more game-like than the control group’s setup. Following the explanation, one final question was prompted by the exit survey, Q7: “if you were part of the other group, how do you think you would have performed?”
5. Results and Discussions
In this results and discussions section of the dissertation, we focus on identifying, describing, and contextualizing the individual results for the numerous analyses conducted. We begin by describing the two key assumptions that were made about our data, including outliers and equality of variance between the two samples. Then, we detail the results for each individual hypothesis (H1-5) and include discussions for each pairwise analysis conducted. Lastly, we describe a few additional analyses that were conducted unrelated to the individual hypotheses but that support the findings of the study as a whole. In the following section, 6. Holistic Discussion, we take on a general perspective and combine all of the individual findings from this section to provide a holistic view of the results and findings of our empirical study.

5.1. Assumptions
This subsection covers how and what assumptions were made about the data to facilitate statistical analyses throughout the remainder of the results. One outlier is identified and removed from further consideration in the dataset and the variance is assumed to be unequal between the two samples. Understanding the data at this level prior to diving into deeper analyses is vital to ensure that the correct statistical analyses are performed, such as the use of Welch’s t-test for difference of means between two samples with unequal variances, as opposed to the use of a t-test for two samples with equal variances or the use of a z-test.

In addition to these assumptions, we utilize an alpha value of 0.10 for establishing statistical significance. This alpha value was chosen since the aim of this research is to show the potential of gamification and to act as a first step towards understanding the effects of gamification on the various items considered in this study (cycle time and perceptions/feelings). At this stage, a lower alpha value is not necessary nor appropriate.

5.1.1. Outliers
An outlier is defined by the National Institute of Standards and Technology (NIST) as, “an observation that lies an abnormal distance from other values in a random sample.” First, using a strip chart of the first attempt cycle time (CT1) data, we were able to identify an outlier within the gamified treatment group.

---
3 Source: 7.1.6. What are outliers in the data? (nist.gov)
Figure 24. Cycle time 1 (CT1) in seconds, separated by treatment group with indication of an outlier in the data set.

The identified outlier depicted in Figure 24 was an outlier for both the gamification participants and the entire data set as whole. Based on review of the observational notes made during the data collection phase, this participant clearly showed signs of struggling with understanding how to build the Lego set from the beginning and did not utilize the provided gamification interface during the first build. Given the extreme nature of this participant, we concluded to remove the outlier for the duration of the analyses. Additionally, all t-tests for hypotheses one, two, and three, outlined below, were performed both with and without the outlier. A comparison of the resulting p-values shows little impact from the removal of the outlier, with no change in results which were identified as significant. The removal of the outlier provides a better fit of the data to expectations provided by the data. Additionally, this brings our data set closer to a 50-50 split of participants, now with nine participants in the control group and ten participants in the gamified group.

5.1.2. Standard Deviations
Next, we examine the standard deviations of the two populations for the first attempt cycle time (CT1) and one representative of the “over time” variable of interest, difference between first and last cycle time (Dif_CT1.CT15). We see that the gamification group, for both variables, results with a much smaller standard deviation value (Table 15). This smaller variability is also visible in the strip charts for both variables (Figure 25 and Figure 26), where we see that the data points for the gamified group results are much closer together with a smaller range compared to the control group data points.
Table 15. Standard deviation for two different variables, to compare the equality of the variance for the control and gamified treatment groups.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Standard deviation of CT1</th>
<th>Standard deviation of Dif_CT1.CT15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1230</td>
<td>1207</td>
</tr>
<tr>
<td>Gamified</td>
<td>427</td>
<td>383</td>
</tr>
</tbody>
</table>

Figure 25. Cycle time 1 (CT1) in seconds by treatment group, for the purpose of analyzing differences in variance between the two samples.
For the purpose of analyzing differences in variance between the two samples.

Thus, moving forward, we assume that the variances are not equal between the two samples based on these factors.

5.2. Hypothesis 1

The first hypothesis (H1) of interest in this study centers on the production performance of the participant at the start of the study. This is formally stated as:

H1: GfM has no effect on the **cycle time** the **first time** an assembly task is completed.

To complete the analysis of this hypothesis, we looked at the first cycle time results (CT1) for the two separate groups, control and gamified. Additionally, we also looked at the cycle time results of the first visit by averaging the first three cycle times (CT1, CT2, CT3) to create an average cycle time for visit 1 (AvgCT_V1) variable to compare between the control and gamification groups. Using R, we looked at the distribution of the variables as a whole, compiled a strip chart to compare between the two groups visually, and then conducted a Welch’s t-test for difference of mean with unequal variances.
From the histogram (Figure 27), we see that the first attempt cycle time (CT1) data is right skewed, with majority of participants having a cycle time between 1500-2500 seconds. We also see that three participants had times in excess of 3500 seconds, with a gap of at least 1000 seconds from other participants.

Figure 28. Strip chart depicting cycle time 1 (CT1) in seconds by treatment group, for the purpose of identifying a difference in means between the two groups.
Next, we use a strip chart (Figure 28) to further analyze the results. We see from the strip chart that the data points for the gamified group are all encompassed within the same range as the control group, showing that the two data sets overlap each other. From looking at the data visually, while the gamification group has a smaller variance than the control group, the means appear to be similar to one another. From this, we suspect that we cannot conclusively state the means of the two samples are significantly different.

Finally, we conduct a Welch’s two sample t-test to assess if the means of the two samples are equal to one another. The t-test results with a p-value of 0.26. Given a desired alpha level of 0.10, we conclude that we cannot reject the null hypothesis. Therefore, we cannot definitively state that the means are different. Thus, we see that the two analyses, the visual approach and t-test approach, support a similar conclusion.

This analysis was also conducted using the average cycle time for the first visit (AvgCT_V1) which was determined by averaging the first three cycle times (CT1, CT2, CT3) for each participant.

![Figure 29. Histogram of average cycle time during visit 1 (AvgCT_V1) in seconds.](image)

The resulting histogram (Figure 29) of the average cycle time for participants for the first visit (first three attempts) (AvgCT_V1) shows a more normal distribution than for the first cycle time (CT1), with the most frequent time being from 1250-1750 seconds. Although this distribution is also right skewed, there appears to be no outliers in this case.
Figure 30. Strip chart depicting average cycle time from visit 1 (AvgCT_V1) in seconds by treatment group.

Using the strip chart (Figure 30), we see that the range of the average first visit cycle time (AvgCT_V1) for the two data sets are more comparable to one another than for the first cycle time (CT1) alone.

In this case, the Welch’s two sample t-test results with a p-value of 0.45. Again, we cannot reject the null hypothesis given the desired alpha level of 0.10. Thus, we conclude that the two means may or may not be equal to one another.

In either case, in absolute terms, the mean for gamification was lower than the mean time for the control group (Table 16).

Table 16. Control group mean, gamified group mean, and resulting p-value of Welch’s t-test for variables related to hypothesis 1 (H1). P-values of significance are highlighted.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Control group mean</th>
<th>Gamified group mean</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>First attempt CT (CT1)</td>
<td>2703.3</td>
<td>2186.6</td>
<td>0.26</td>
</tr>
<tr>
<td>Average first visit CT (AvgCT_V1)</td>
<td>2041.9</td>
<td>1822.1</td>
<td>0.45</td>
</tr>
</tbody>
</table>

While we cannot definitively state that the means are different from one another, we can see a trend in the data that shows gamification tends to result in a lower cycle time early in the case of a repetitive, monotonous assembly task.

5.3. Hypothesis 2
The second hypothesis (H2) of interest in this study centers on the production performance of the participant over a period of time. This is formally stated as:

H2: GfM has no effect on cycle time of an assembly task over time (5 days, 2-3 hours per day).
To complete the analysis of this hypothesis, we considered the varying sets of cycle time results for the two separate groups, control and gamification. First, we looked at the difference between the first and the last cycle time (Dif_CT1.CT15). Second, we looked at the difference between the cycle times for the first and last visit, by taking the difference of the average of the first three cycle times and the average of the last three cycle times (Dif_AvgCTV1.AvgCTV5). Third, we looked solely at the final cycle time (CT15). Last, we looked at the cycle time for the last visit, by averaging the last three cycle times (AvgCT_V5). Using R, for each of the four aforementioned variables, we looked at the distribution of the variables as a whole, compiled a strip chart to compare between the two groups visually, and then conducted a Welch’s t-test for difference of means with unequal variances.

5.3.1. Difference between First and Last Cycle Time
First, we look at the distribution of the data using a histogram which shows a right-skewed data set.

![Figure 31. Histogram of difference from cycle time 1 to cycle time 15 (Dif_CT1.CT15) in seconds.](image)

Similar to the histogram (Figure 27Figure 31) for the first cycle time (CT1) data (Hypothesis 1), we see a set of points that are disjointed from the rest of the data set, with a difference from first to last cycle time (Dif_CT1.CT15) of more than 2750 seconds (Figure 31).

Then, we use a strip chart to visually showcase the data, split between the two groups.
Figure 32. Strip chart depicting difference from cycle time 1 to cycle time 15 (Dif_CT1.CT15) in seconds by treatment group, for the purpose of identifying a difference in means between the two groups.

From the strip chart (Figure 32), we see that the gamification participants have a much smaller variance and range than the control group participants.

Last, we perform a Welch’s t-test which assumes that the two populations have different variances. The result shows no significant difference for the difference from first to last attempt cycle times, with a p-value of 0.3579.

5.3.2. Difference between First Visit and Last Visit Cycles Times
Similar to the above analysis, we analyzed the results for the difference between the first visit average cycle times (AvgCT_V1) and last visit average cycle times (AvgCT_V5).
Figure 33. Histogram of difference between the first visit average cycle times and last visit average cycle times (Dif_AvgCTV1.AvgCTV5) in seconds.

The resulting histogram (Figure 33) is on a much smaller range than the previous histogram that looked at the difference between first and last attempts (Dif_CT1.CT15) only. Again, we see a distribution of times resulting in a right-skewed histogram (Figure 31).

Figure 34. Strip chart depicting difference between the first visit average cycle times and last visit average cycle times (Dif_AvgCTV1.AvgCTV5) in seconds by treatment group.
The strip chart (Figure 34) generates a similar view of the data that we have seen so far, with similar ranges for the two samples.

Ultimately, the Welch’s two sample t-test results with a p-value of 0.5813 and again we cannot reject our null hypothesis.

The difference between cycle times, whether for the first to last attempt (Dif_CT1.CT15) or the first to last visit (Dif_AvgCTV1.AvgCTV5), is a representation of how much learning and improvement each participant made during the duration of the study. A larger difference indicates that the participant improved more than those with lower values. For the purpose of examining the differences from start to finish of the study, although these time improvements do not provide statistically significant results, we assume that participants had less room for improvement by the conclusion of the study, since participants in the gamified group on average had a lower first attempt cycle time (CT1) and first visit average cycle time (AvgCT_V1). Therefore, measuring the performance of participants later into the study can be better understood by analysis of the final attempt cycle time (CT15) and final visit cycle time (AvgCT_V5).

5.3.3. Final Attempt Cycle Time

In an interest to understand whether gamification has any effect on the participants after they have been working with it for a period of time, we looked solely at the final cycle time (CT15) for the two groups, control and gamified.

The distribution of the times for the final attempt is not as right-skewed as previous variables but does still show two instances of seemingly abnormally high final cycle times (CT15) (Figure 35). Conversely to previous variable distributions, the distribution of CT15 appears to have a median and mean that falls near one another and within the most common category, indicating that the distribution is more normal with less skewness.
Between the two groups, we see that the control group participants have two instances of cycle times which exceed the maximum cycle time for gamification participants (Figure 36). Additionally, the gamification group has approximately five cycle times which are less than the minimum cycle time for the control group. The range of the two data sets appears to be comparable, and the data appear to be spread similarly between the groups. Ultimately, the chart indicates that gamification participants may show a final cycle time that is on average less than the control group.

With this inclination, we perform a Welch’s two sample t-test for difference of means which results in a p-value of 0.066. At a desired alpha level of 0.10, we determine that we reject the null hypothesis in this case. Therefore, we conclude that GfM results in a lower cycle time for the final attempt (CT15) of a manufacturing assembly task in comparison to the control.

Observations from the experiment brought to light that this may have been caused by participants’ last-ditch efforts to obtain a new personal record badge. We believe that participants wanted to end the study on a high note by showcasing their best efforts. This makes us wonder if participants would have had such a low final cycle time given that they were not aware this was the final attempt.

However, we see that participants in both groups showed efforts in improving on the last attempt. Out of the nine control participants, seven (approx. 78%) showed improvement from the 14th to 15th attempt. Out of the ten gamification participants, nine (90%) showed improvement from the 14th to 15th attempt. It is difficult to say if the gamification caused more participants to improve on the last attempt, or if this was a drive for all participants in the study, regardless of treatment group.

In either case, we see that gamification participants complete the study with a lower cycle time than control participants on average. Therefore, this portion of the analysis supports the conclusion that we...
reject our second hypothesis (H2), which states that GfM has no effect on cycle time of an assembly task over time.

5.3.4. Final Visit Cycle Time
Next, we looked at the average of the final three attempt cycle times to review the average cycle time of the last visit (AvgCT_V5). This was expected to mitigate the effects of any last-ditch efforts of participants to showcase their best abilities since this was over three trials, not just the very last attempt.

The histogram (Figure 37) shows that the data points are again more right skewed, with majority of participants having a final visit average cycle time between 750 and 850 seconds.

Figure 37. Histogram of the average cycle time of the last visit (AvgCT_V5) in seconds.

Figure 38. Strip chart depicting average cycle time for visit 5 (AvgCT_V5) in seconds by treatment group.
The strip chart (Figure 38) shows significant overlapping of data points; however, the gamified group data points appear to average less than the control data points. Again, the range and variation of points appear to be similar between the groups.

The resulting p-value of the test is 0.22, indicating that we cannot reject our null hypothesis at a desired alpha level of 0.10.

Considering the average cycle time for the last visit (AvgCT_V5) against the last attempt cycle time (CT15), since the gamification participants do not outperform the control participants as significantly when considering the average last visit cycle time (AvgCT_V5), we believe that this speaks to the argument that improvement in cycle time performance due to last-ditch efforts were more significant for gamified group participants than for control group participants.

In any case, the mean for gamification was again lower than the mean time for the control group (Table 17).

Table 17. Control group mean, gamified group mean, and resulting p-value of Welch’s t-test for variables related to hypothesis 2 (H2). P-values of significance are highlighted.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Control group mean</th>
<th>Gamified group mean</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference between First and Last CT (CT1-CT15)</td>
<td>1798.9</td>
<td>1393.8</td>
<td>0.36</td>
</tr>
<tr>
<td>Difference between First and Last Visit CT (Avg(CT1:CT3)-Avg(CT13:CT15))</td>
<td>1094.6</td>
<td>946.4</td>
<td>0.58</td>
</tr>
<tr>
<td>Last Attempt CT (CT15)</td>
<td>904.4</td>
<td>792.8</td>
<td>0.066</td>
</tr>
<tr>
<td>Average Last Visit CT (Avg CT13:CT15)</td>
<td>947.3</td>
<td>875.7</td>
<td>0.22</td>
</tr>
</tbody>
</table>

While we cannot definitively state that the means are different from one another in three out of four of the analyses for this hypothesis, we can see a trend in the data that shows gamification tends to result in a lower cycle time after a period of time in the case of a repetitive, monotonous assembly task.

In conclusion, our experiment can reject our second hypothesis (effect of GfM on cycle time over time) in one instance (CT15). While other instances of testing this hypothesis do not have statistical significance, they still show that gamification tends to result in lower cycle times on average.

The difference in cycle times from the first and last attempt (Dif_CT1.CT15) and the first and last visit (Dif_AvgCTV1.AvgCTV5) shows how much the participants improved over the course of the study. Although this theoretically shows that gamification participants improved less than control participants, coupled with the average lower first (CT1) and last cycle times (CT15), these lesser improvement values might be attributed to the fact that there was less room or opportunity for improvement for participants in the gamified treatment group from the start. Therefore, this may indicate that gamification supported participants in performing better from the very first instance of completing the assembly task.
5.4. Hypothesis 3

The third hypothesis (H3) of interest in this study centers on the psychological effects of gamification on the participant at the start of the study. This is formally stated as:

H3: GfM has no effect on the perceptions/feelings of users the first time an assembly task is completed.

To complete the analysis of this hypothesis, we looked at the results of the NASA TLX questionnaire on the first visit for the two separate groups, control and gamified. The NASA TLX questionnaire results with six data points for each user, one for each of the following aspects: mental demand, physical demand, temporal demand, performance, effort, and frustration. The app\(^4\) used for the NASA TLX questionnaire, published by NASA for use on Apple devices, defines the six aforementioned categories as:

- **Mental demand (low/high):** How much mental and perceptual activity was required (for example, thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, forgiving or exacting?
- **Physical demand (low/high):** How much physical activity was required (for example, pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
- **Temporal demand (low/high):** How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
- **Performance (good/poor):** How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
- **Effort (low/high):** How hard did you have to work (mentally and physically) to accomplish your level of performance?
- **Frustration level (low/high):** How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

Using R, we looked at the distribution of the variables as a whole, compiled a strip chart to compare between the two groups visually, and then conducted a Welch’s t-test for difference of mean with unequal variances.

5.4.1. NASA TLX Scores from Visit 1

In this section, we take a look at the six NASA TLX scores from the very first visit.

5.4.1.1. Mental Demand from Visit 1

The mental demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” mental demand and 100 being “high” mental demand. Mental demand is measured by answering the question: “how much mental and perceptual activity did you spend for this task?”

\(^4\) App name: NASA TLX, Developer: NASA, Version: 1.0.3, NASA Point of Contact: Brian Gore (brian.f.gore@nasa.gov), Obtained via: Apple App Store, For use on: Apple iPad Mini (6\(^{th}\) Generation)
Reviewing the histogram of the data (Figure 39), we see one outlier who rated the task as having very low mental demand. Otherwise, the ratings are mostly scattered at values in excess of 25.

From the strip chart (Figure 40), we see that the participant who rated the experiment as having no mental demand was part of the control group and indicated “0” mental demand required. Aside from this
participant, we see that the ratings between the two groups are very similar in both mean, spread, and range.

Ultimately, the Welch’s t-test returns a p-value of 0.92 indicating that the two means are not different from one another. The control group had a mean mental demand rating of 54.4 and the gamification group had a mean mental demand rating of 55.5.

This is perceived as positive for gamification because it indicates that gamification does not cause additional mental workload for participants when they are beginning to learn a new process. While gamification does not work to mitigate some of the mental demand, it does not cause excess load on individuals. This is a common concern in the field: that having the additional task of using and understanding the gamification interface in addition to complete the task at hand can lead to distraction and overload for workers.

5.4.1.2. Physical Demand from Visit 1

The physical demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” physical demand and 100 being “high” physical demand. Physical demand is measured by answering the question: “how much physical activity did you spend for this task?”

![Histogram of NASA TLX Physical Demand scores from visit 1 (TLX.PD.V1).](image)

The resulting histogram (Figure 41) for physical demand on the first visit shows an inverted bell curve, with a bimodal distribution with modes occurring at 5-15 and 55-65. The upper limit of physical demand is much lower than the upper limit of the mental demand ratings for the first visit. This indicates that the experimental setup is likely to be more mentally demanding than physically demanding. This was intentional in the design of the experiment as we were most interested in seeing how gamification impacts the psychological aspects of simple, monotonous assembly tasks.
From the strip chart (Figure 42), we see that the control participants mostly rated physical demand low, with a few outliers. Meanwhile, gamification participants rated physical demand at varying levels.

The Welch’s t-test returns a p-value of 0.39, indicating that we cannot reject our null hypothesis that physical demand is affected by the application of gamification. The resulting means for the control group and gamification group are 23.3 and 32.5, respectively.

Although there is no statistical difference in means here, there is quite a difference in the actual mean values that result from the two groups. We are not able to hypothesize many explanations for a difference in the means here, given that the (physical) tasks were the exact same. A possible explanation that comes to mind is that participants may have potentially mixed their feelings of mental demand with physical demand for the study. However, this is not based on observation during the experiments and needs to be verified in future research.

5.4.1.3. Temporal Demand from Visit 1

The temporal demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” temporal demand and 100 being “high” temporal demand. Temporal demand is measured by answering the question: “how much time pressure did you feel in order to complete this task?”
From the histogram (Figure 43), we see that the mode occurs in the 75-85 range. The ratings are spread almost entirely across all possible values, with a majority occurring above a value of 50.

Figure 44. Strip chart depicting NASA TLX Temporal Demand scores from visit 1 (TLX.TD.V1) by treatment group.

The strip chart (Figure 44) of the temporal demand rated for the two groups shows that control participants tend to rate their temporal demand lesser than gamification participants.
The Welch’s t-test returns a p-value of 0.076, indicating that we can reject the null hypothesis and conclude that gamification causes an increase in temporal demand and sense of time pressure to complete a task. The resulting means for the two groups are 45.0 for the control group and 69.0 for the gamification group.

This finding compared to the result that gamification participants showed an average lower cycle time for the first attempt and first visit indicates that the added (perceived) time pressure can be a positive influence on participants to perform better (measured by productivity metrics like cycle time) when gamification is added to a manufacturing task. Although the temporal demand was higher for gamification participants, this did not cause an increase in mental demand. Thus, indicating the increased sense of time pressure did not negatively impact users’ psychological balance during the study.

5.4.1.4. Performance from Visit 1

The performance is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “good” performance and 100 being “poor” performance. It is important to note the location of the endpoints with respect to their qualitative meanings. This is often the opposite of what would be expected. Performance is measured by answering the question: “how successful do you think you were in accomplishing the goals of the task?”

![Figure 45. Histogram of NASA TLX Performance scores from visit 1 (TLX.P.V1).](image)

The histogram (Figure 45) of performance ratings show that a vast majority of participants rated themselves with “good” performance (ratings less than 25). Only three participants rated themselves with a performance rating above 25. Of those three participants, there are two clear outliers, one occurring at 45-55 and the other at 85-95. As a reminder, the higher performance rating value, the poorer a participant believed their performance to be.
Figure 46. Strip chart depicting NASA TLX Performance scores from visit 1 (TLX.P.V1) by treatment group.

From the strip chart (Figure 46), we see one outlier was from the control group and the most extreme outlier was from the gamification group. Conversely, we see that more individuals in the gamification group rate themselves in the lowest bracket (0-5) compared to the control group participants.

Ultimately, the Welch’s t-test returns a p-value of 0.54 indicating that we cannot reject the hypothesis. The resulting means are 15.6 and 22.0 for the control group and gamification group, respectively.

There does not appear to be any significant trend here. The gamification outlier may have significantly impacted the resulting mean for the gamification group, and similarly for the control group and its respective outlier. Aside from looking at the averages and outliers, we can observe that individuals between the two groups rated themselves fairly similarly in the 0 to 25 range.

From the analysis, we see that the two samples may have similar means to one another. There are many interpretations for understanding the significance of this finding. In one fashion, this is a positive for our study as we had hoped to create an equal experiment for participants in both cases. Given that participants perceived their performance similarly to those in the other treatment group, we conclude that participants were equally confident in both groups with how they performed. Another potential interpretation of this result is that, in support of GfM, it may have been more promising to see a lower (better) performance score for the gamified group since they received feedback on their builds, and therefore would have been more confident in their builds. Conversely to this idea, we may have expected control participants to rate their performance lower (better) because, without feedback, they would assume their efforts were positive. This is based on the idea “ignorance is bliss,” which essentially means that not knowing something may lead to a more positive experience, or in other words, you cannot worry about the things you do not know. Regardless of the interpretation or our mind wanderings discussed herein, more research is needed to gather a full understanding of how gamification may impact perceived performance. This is discussed with additional recommendations in the 7.2. Limitations and Future Work.
5.4.1.5. Effort from Visit 1
The effort is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” effort and 100 being “high” effort. Effort is measured by answering the question: “how hard did you have to work to accomplish your level of performance?”

The resulting histogram (Figure 47) for effort scores shows that participants primarily judged their effort at about average, with the mode occurring right at the center rating of 45-55.

Figure 48. Strip chart depicting NASA TLX Effort scores from visit 1 (TLX.E.V1) by treatment group.
The strip chart (Figure 48) does not indicate much, except that control participants rated their effort on a wider range. It is possible that the two outliers of the control group with scores of 75 and 80 are individuals who had to restart or go back significantly on their build progress due to inaccuracies in their build. This appeared to have significant impact on control group participants since they were not provided feedback on their build by the researcher until the build was believed to be completed. Often, this involved significant and unguided backtracking to locate and resolve the mistake. For gamification participants, they were provided clear and timely information of where their error had occurred. Additionally, the feedback to gamification participants was continuous throughout the study, as long as they referred to the screen. Thus, they tended to not have to backtrack as many steps because they often would catch their mistake only a few steps later, as opposed to all the way at the perceived completion of the build, when the researcher would alert them of an inaccuracy.

Ultimately, the Welch’s t-test resulted with a p-value of 0.29, indicating that we cannot reject the null hypothesis and we cannot conclude the means of the two groups are different. The resulting means for the control group and gamification group are 44.4 and 53.5, respectively.

For effort ratings, it is interesting to see that the average rating was higher for gamification participants than control participants, even if not statistically significant. Given the lack of feedback and strict requirement that the build be completed correctly for all participants, we anticipated that control group participants would have a higher effort score. Conversely, we hypothesize that gamification participants resulted with a higher effort score since they had feedback which informed them of their performance throughout. With increased feedback and seemingly no room for error (gamification interface appeared to keep track of accuracy at every step), gamification participants may have put in more effort to their builds, resulting in the higher average effort score for gamification participants compared to control participants.

5.4.1.6. Frustration from Visit 1

The frustration is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” frustration and 100 being “high” frustration. Frustration is measured by answering the question: “how insecure, discouraged, irritated, stressed, and annoyed were you during this task?”
Frustration scores range the entire span of results, from 0-5 all the way up to 95-100 (Figure 49). The ratings appear to be mostly spread out across the potential values, with a mode occurring at a rating of 0-5.

The strip chart (Figure 50) shows that gamification participants have two clusters, one below 10 and one above 50, while majority of the control participants responded with a rating below 50.
The Welch’s t-test results with a p-value of 0.45 indicating that we cannot reject our hypothesis and that the means are not statistically different from one another. The resulting means are 33.9 and 45.0 for the control group and gamification group, respectively.

Similar to the effort scores, we are surprised by the results here. In one case, we expected gamification participants to have a lower frustration score since we expected them to enjoy the experience more than the control participants. However, the frustration in this case may be positive in its effects of improved accuracy and therefore performance. Additionally, the lower control group mean may indicate that these participants were disconnected from the situation and did not experience significant emotion, either positively or negatively. In another consideration, we had expected the control group participants to experience higher frustration given the lack of feedback and necessity to return to an unknown position to resolve errors, especially in the early runs of the study. However, it is possible that during the first run, participants in all cases were more careful with their building, so fewer mistakes were made and therefore, less frustration occurred.

### 5.4.1.7. Conclusions on NASA TLX Scores from Visit 1

In conclusion, we obtained the following results for H3 (Table 18):

Table 18. Control group mean, gamified group mean, and resulting p-value of Welch’s t-test for variables related to hypothesis 3 (H3). P-values of significance are highlighted.

<table>
<thead>
<tr>
<th>Variable (score at visit 1)</th>
<th>Control group mean</th>
<th>Gamified group mean</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Demand</td>
<td>54.4</td>
<td>55.5</td>
<td>0.92</td>
</tr>
<tr>
<td>Physical Demand</td>
<td>23.3</td>
<td>32.5</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>Temporal Demand</strong></td>
<td><strong>45.0</strong></td>
<td><strong>69.0</strong></td>
<td><strong>0.076</strong></td>
</tr>
<tr>
<td>Performance</td>
<td>15.6</td>
<td>22.0</td>
<td>0.54</td>
</tr>
<tr>
<td>Effort</td>
<td>44.4</td>
<td>53.5</td>
<td>0.29</td>
</tr>
<tr>
<td>Frustration</td>
<td>33.9</td>
<td>45.0</td>
<td>0.45</td>
</tr>
</tbody>
</table>

### 5.5. Hypothesis 4

The fourth hypothesis (H4) of interest in this study centers on the psychological effects of gamification on the participant over a period of time. This is formally stated as:

**H4:** GfM has no effect on the perceptions/feelings of users for an assembly task over time.

To complete the analysis of this hypothesis, we again utilized the NASA TLX questionnaire. Similar to hypothesis 2 (H2) that looked at the cycle time changes over time, in this analysis, we look at the NASA TLX assessment scores for i) the difference between the first and last (fifth) visit, ii) the last visit, and iii) the average of all five visits.

As mentioned for hypothesis 3 (H3), the NASA TLX questionnaire results with six data points, one for each of the following aspects: mental demand, physical demand, temporal demand, performance, effort, and frustration. This data was collected using the previously introduced NASA TLX app.

Using R, we looked at the distribution of the variables as a whole, compiled a strip chart to compare between the two groups visually, and then conducted a Welch’s t-test for difference of mean with unequal variances.
5.5.1. NASA TLX Scores Difference from Visit 1 to Visit 5

In this section, we take a look at the difference of the six NASA TLX scores from the first visit to the last visit. To find the difference of each category score from the first visit to the last visit (fifth), we subtracted the score of visit five from the score of visit one for each participant (Score_V1-Score_V5). There was potential for scores at visit five to be higher than at visit one, indicating that a participant’s perceived workload was higher later in the study, resulting in a negative value for this analysis.

5.5.1.1. Mental Demand Difference from Visit 1 to Visit 5

The mental demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” mental demand and 100 being “high” mental demand. Mental demand is measured by answering the question: “how much mental and perceptual activity did you spend for this task?” To find the difference of the mental demand from visit one to visit five, we subtracted the mental demand score of visit five from the mental demand score of visit one for each participant (MD_V1-MD_V5). There was potential for mental demand scores at visit five to be higher than visit one, indicating that a participant’s mental demand was higher later in the study, resulting in a negative value for this analysis.

![Figure 51. Histogram of difference of the mental demand scores from visit one to visit five (Dif_MDV1.MDV5).](image)

The resulting histogram (Figure 51) for all difference of mental demand scores from visit one to visit five (Dif_MDV1.MDV5) shows a wide spread of responses, ranging from well below zero to nearly 75. We see that the resulting values for the difference are spread from significant improvement (higher values) to mild retrogression (values below zero).
Figure 52. Strip chart depicting difference of the mental demand scores from visit one to visit five (Dif_MDV1.MDV5) by treatment group.

From the strip chart (Figure 52) of the data, we see a clear difference between the control and gamification groups. The control participants clearly have a higher improvement in their mental demand scores in comparison to gamification participants. The control group has two outliers at or below 0, while the gamified group has four participants at or below 0, all of which are non-outliers to the gamified group. The gamified group has one outlier near the value of 60, while the control group appears to have a cluster of results around this value.

The Welch’s t-test results with a p-value of 0.042. The mean for the control group is 40.0 and the mean for the gamified group is 14.5. Therefore, we conclude that the control group has a significantly higher difference in mental demand scores from the first to last visit, indicating a lesser mental demand at the conclusion of the study, in comparison to the participants in the gamified group.

Initially, we hypothesized that this showed gamification to be a poor tool to use because it caused higher mental demand for participants later in the study in comparison to the control group, and because in four out of ten (40%) participants, their mental demand actually increased by the end of the study compared to their initial score at the first visit. We thought this would cause increased mental stress and load on the participants, however, after further deliberation we concluded that mental demand in a monotonous task is not necessarily problematic. In fact, increased mental demand helps participants stay actively involved in and engaged with the task at hand. In essence, gamification may act as a tool for countering the boreout\(^5\) challenge of repetitive tasks. Especially in conjunction with the improved cycle times seen for

---

\(^5\) “boreout” is a syndrome resulting from tasks which are perceived as boring or monotonous. The lack of mental stimulation can have a detrimental effect on the mental state (and thus productivity) of those involved in completing such tasks. The effects of “boreout” are similar to the effects of “burnout,” however, the cause of these two syndromes are opposite of one another (boredom and over exertion, respectively). (Source: Why Boredom At Work Is More Dangerous Than Burnout (forbes.com))
gamification participants, we believe that the increased mental demand later in the study for gamification participants (in comparison to control participants) is a positive impact of gamification for monotonous assembly tasks. Gamification provides its users with a reason and a method to stay actively engaged with the task, leading to improved learning and quality, ultimately resulting in increased production efficiency.

5.5.1.2. Physical Demand Difference from Visit 1 to Visit 5
The physical demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” physical demand and 100 being “high” physical demand. Physical demand is measured by answering the question: “how much physical activity did you spend for this task?” To find the difference of the physical demand from visit one to visit five, we subtracted the physical demand score of visit five from the physical demand score of visit one for each participant (PD_V1-PD_V5). There was potential for physical demand scores at visit five to be higher than visit one, indicating that a participant’s physical demand was higher later in the study, resulting in a negative value for this analysis.

Figure 53. Histogram of difference of the physical demand scores from visit one to visit five (Dif_PDV1.PDV5).

The resulting histogram (Figure 53) of the values for the difference in physical demand from visit one to visit five shows an overwhelming majority indicating 0 change in physical demand from the start to finish of the study. This is promising because this is what we hoped to see since the physical aspects of the study did not change whatsoever throughout the duration of the study.
Figure 54. Strip chart depicting difference of the physical demand scores from visit one to visit five (Dif_PDV1.PDV5) by treatment group.

From the strip chart (Figure 54), we see one upper and one lower outlier for each group, with majority of participants in each group having results around the -5 to 20 range. There appears to be no significant difference between the two groups here.

Ultimately, the Welch’s t-test results with a p-value of 0.48 indicating no statistical difference in means between the two groups. The resulting mean for the control group is 9.4, while the resulting mean for the gamified group is 3.5.

Since both groups completed the exact same physical tasks, it is promising to see a result indicating similar physical demand i) throughout the study and ii) between the two groups. Although there is no statistical difference in the means, we see that the gamified group has a lower mean than the control group. This may be related to the increased mental demand experienced by gamification participants, thereby leading the participants in the gamified group to pay less attention to the physical demands of the study. Additionally, it is possible that the control participants have no outside factors to focus on aside from the physical task, and therefore, they are more aware of the effects of the physical task especially as the study progresses and they become increasingly bored of the task at hand.

5.5.1.3. Temporal Demand Difference from Visit 1 to Visit 5

The temporal demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” temporal demand and 100 being “high” temporal demand. Temporal demand is measured by answering the question: “how much time pressure did you feel in order to complete this task?” To find the difference of the temporal demand from visit one to visit five, we subtracted the temporal demand score of visit five from the temporal demand score of visit one for each participant (TD_V1-TD_V5). There was potential for temporal demand scores at visit five to be higher than visit one, indicating that a participant’s temporal demand was higher later in the study, resulting in a negative value for this analysis.
Again, we see a varying degree of results for difference of temporal demand scores from visit one to visit five (Dif_TDV1.TDV5) ranging from well below 0 to upwards of 75 (Figure 55). There does not appear to be any significant results of note, and many responses are spread uniformly, with no more than three participants in any given category.

Figure 56. Strip chart depicting difference of the temporal demand scores from visit one to visit five (Dif_TDV1.TDV5) by treatment group.
Similar to the difference of physical demand scores from visit one to visit five (Dif_PDV1.PDV5) results (Figure 54), we see very little difference between the two groups here (Figure 56). Most notably, we see that the control group has a wider range of results, from -25 to 70, while the gamified group ranges only from -5 to 60.

The Welch’s t-test results with a p-value of 0.56. The resulting means are 29.4 and 21.5 for the control group and gamified group, respectively.

Since temporal demand is with regards to how much time pressure each participant felt, we would expect the two groups to show similar improvements from the beginning to the end of the study since there were no changes within each group. For both groups, we would expect to see similar changes in temporal demand because the task and treatment are not changing from the start to the finish of the study. Ultimately, we see that both groups experience a slight improvement in temporal demand, on average, indicating that they felt slightly less time pressure from the start to finish of the study in each group.

5.5.1.4. Performance Difference from Visit 1 to Visit 5

The performance is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “good” performance and 100 being “poor” performance. It is important to note the location of the endpoints with respect to their qualitative meanings. This is often the opposite of what would be expected. Performance is measured by answering the question: “how successful do you think you were in accomplishing the goals of the task?” To find the difference of the participant’s perceived performance from visit one to visit five, we subtracted the performance score of visit five from the performance score of visit one for each participant (P_V1-P_V5). There was potential for performance scores at visit five to be higher than visit one, indicating that a participant’s perceived performance was higher (worse) later in the study, resulting in a negative value for this analysis.

![Histogram of difference of the performance scores from visit one to visit five (Dif_PV1.PV5).](image)

Figure 57. Histogram of difference of the performance scores from visit one to visit five (Dif_PV1.PV5).
The resulting histogram (Figure 57) shows that the vast majority of participants rated themselves with nearly the same initial and final performance scores, with majority of difference of performance scores from visit one to visit five (Dif_PV1.PV5) falling in the -5 to 5 range. There are two outliers in the data set, one in the 35-45 block and one in the 85-95 block.

![Figure 58. Strip chart depicting difference of the performance scores from visit one to visit five (Dif_PV1.PV5) by treatment group.](image)

From the strip chart (Figure 58), which separates the data into their respective treatment groups, we see that one of the outliers was a difference in scores of 45 for the control group, while the more extreme outlier (value = 85) resided with the gamified group. Ultimately, it appears as though the two groups had roughly similar results when looking at the non-outliers.

The resulting p-value from the Welch’s t-test for difference of means is 0.90 indicating no statistical difference between the two groups. The resulting means are 8.9 and 7.5 respectively.

The results of this test indicate that, regardless of which treatment group, participants saw approximately the same change in their performance from the start to finish of the study. Interestingly, the performance scores only improved on an average of roughly eight points, which is comparatively small considering some of the other average scores seen. In either case, participants rated themselves with minor performance improvements, on average. Gamification had no impact on how participants perceived the change in their performance from start to finish of the study.

5.5.1.5. Effort Difference from Visit 1 to Visit 5

The effort is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” effort and 100 being “high” effort. Effort is measured by answering the question: “how hard did you have to work to accomplish your level of performance?” To find the difference of the participant’s perceived effort from visit one to visit five, we subtracted the effort score of visit five from the effort score of visit one for each participant (E_V1-E_V5). There was potential for effort scores at visit five to be higher than visit one,
indicating that a participant’s perceived effort was higher later in the study, resulting in a negative value for this analysis.

![Histogram of difference of the effort scores from visit one to visit five (Dif_EV1.EV5).](image)

Figure 59. Histogram of difference of the effort scores from visit one to visit five (Dif_EV1.EV5).

The resulting histogram (Figure 59) for the difference in effort scores (Dif_EV1.EV5) shows again a wide range of values. Most notably, we see a significantly negative result in the range of -45 to -55, indicating one participant had significantly higher efforts later in the study. The scores vary heavily between this extreme negative value, all the way up to a range of 55 to 65. The most common occurrence is a difference in the range of -5 to +5, however this is not a significant mode value as there are other values with similar levels of occurrence (-15 to -5 and 45 to 55).

![Strip chart difference of effort scores from visit 1 to visit 5 (Dif_EV1.EV5) by treatment group.](image)

Figure 60. Strip chart difference of effort scores from visit 1 to visit 5 (Dif_EV1.EV5) by treatment group.
From the strip chart (Figure 60), we see that the extreme negative instance was from a gamified group participant. Similarly, the two most extreme data points are from the gamified group. The gamified group shows more variance than the control group for these ratings. The control group has values ranging from -10 to +55, while the gamified group has values ranging from -50 to +60.

The Welch’s two sample t-test for difference of means results with a p-value of 0.19. The control group mean is 27.2 and the gamified group mean is 8.0. Based on the resulting p-value, we conclude there is no statistical difference in means for the two groups.

If we remove the outliers, visually speaking, it appears as though the two groups may result with statistically different means, with the gamified group having a lower mean. Aside from this hypothesis, we also see that the gamified group mean is lower than the control group mean, even though it is not statistically different. We hypothesize that this is similar to the rationale of the lesser improvement of the mental demand scores for the gamified group as discussed in 5.5.1.1. Mental Demand Difference from Visit 1 to Visit 5. We believe that the gamified participants had, on average, less improvement (lower Dif_EV1.EV5 values) from the start to finish of the study because they remained actively engaged throughout the study in an attempt to improve their productivity. Additionally, we believe that gamification directly kept the participants mentally engaged throughout the study by giving them something (the gamification interface screen) to check in with to ensure quality/accuracy of their build.

5.5.1.6. Frustration Difference from Visit 1 to Visit 5
The frustration is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” frustration and 100 being “high” frustration. Frustration is measured by answering the question: “how insecure, discouraged, irritated, stressed, and annoyed were you during this task?” To find the difference of the participant’s frustration from visit one to visit five, we subtracted the frustration score of visit five from the frustration score of visit one for each participant (F_V1-F_V5). There was potential for frustration scores at visit five to be higher than visit one, indicating that a participant’s frustration was higher later in the study, resulting in a negative value for this analysis.

Figure 61. Histogram of difference of the frustration scores from visit one to visit five (Dif_FV1.FV5).
The histogram (Figure 61) of the difference in frustration scores (Dif_FV1.FV5) shows three separate clusters, one at the far negative spectrum, one with little to no change, and another with moderate to significant improvements.

![Histogram](image)

Figure 62. Strip chart depicting difference of the frustration scores from visit one to visit five (Dif_FV1.FV5) by treatment group.

When looking at the strip chart (Figure 62) which breaks the two treatment groups apart, we see that the two groups appear to have fairly similar data, aside from the one extreme negative data point within the gamified treatment group. No conclusions can be drawn from this chart.

The Welch's t-test results with a p-value of 0.30. The control group mean is 26.7 and the gamified group mean is 11.0. The conclusion from the Welch's t-test is that the difference between these means is not significant, therefore, we cannot conclusively state that either treatment impacts the improvement of frustration over the course of the study.

By solely looking at the mean values, we see that the control group had a higher mean, indicating an improvement of frustration scores throughout the duration of the study in comparison to the average of the gamified group. Similarly to the effort and mental demand scores, we believe this is not necessarily a bad attribute of gamification, but instead keeps participants motivated and actively engaged throughout the build process. This is particularly important for monotonous assembly tasks to improve quality and productivity. Although higher scores of frustration are intuitively less appealing, considering the evolution-based idea that some stress is the ideal state, we believe that the moderate amounts of frustration brought on by gamification is promising for the future of gamification and aligns with its objective.

This (referring to motivated and actively engaged participants) relates to the state of “flow” often discussed in gamification research, in which users enter a state of just flowing through a task, where everything goes smoothly, is mentally and physically engaging, and feels as if it comes naturally. It is believed by many that the flow state can lead to increased efficiency for monotonous tasks. This research
is a testament to this through results that show lower (improved) cycle times and higher NASA TLX scores for participants in the gamified treatment group, indicating better production efficiency and improved mental and physical engagement for a monotonous assembly task.

### 5.5.1.7. Conclusions on Difference in NASA TLX Scores from Visit 1 to Visit 5

In conclusion, we obtained the following results for H4, when looking at the difference of the TLX scores from the first and last visit (Table 19):

Table 19. Control group mean, gamified group mean, and resulting p-value of Welch’s t-test for variables related to difference in NASA TLX Scores from Visit 1 to Visit 5. P-values of significance are highlighted.

<table>
<thead>
<tr>
<th>Variable (difference in scores from visit 1 to visit 5)</th>
<th>Control group mean</th>
<th>Gamified group mean</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Demand</td>
<td>40.0</td>
<td>14.5</td>
<td>0.042</td>
</tr>
<tr>
<td>Physical Demand</td>
<td>9.4</td>
<td>3.5</td>
<td>0.48</td>
</tr>
<tr>
<td>Temporal Demand</td>
<td>29.4</td>
<td>21.5</td>
<td>0.56</td>
</tr>
<tr>
<td>Performance</td>
<td>8.9</td>
<td>7.5</td>
<td>0.90</td>
</tr>
<tr>
<td>Effort</td>
<td>27.2</td>
<td>8.0</td>
<td>0.19</td>
</tr>
<tr>
<td>Frustration</td>
<td>26.7</td>
<td>11.0</td>
<td>0.30</td>
</tr>
</tbody>
</table>

In all cases, we see that the gamified treatment group has a lower mean than the control treatment group. This may be attributed to the idea that gamification expedites learning, thus participants in the gamified group were already settled in their beliefs and understanding of the mock assembly task from early in the study. Therefore, participants in the gamified group had less room for improvement since they learned early in the study how they related to each of the six NASA TLX categories. Whereas, in the control group, they were uncertain about their perceptions of the task in the first iteration, and therefore, rated these variables higher than they otherwise would have after additional iterations of the build.

### 5.5.2. NASA TLX Scores at Visit 5

In this subsection, we take a look at the scores from the six NASA TLX categories from the very last visit.

#### 5.5.2.1. Mental Demand at Visit 5

The mental demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” mental demand and 100 being “high” mental demand. Mental demand is measured by answering the question: “how much mental and perceptual activity did you spend for this task?”
Figure 63. Histogram of NASA TLX mental demand scores from visit five (TLX.MD.V5).

The resulting histogram (Figure 63) shows a generally decreasing frequency as mental demand (TLX.MD.V5) scores increase. Majority of participants rated their mental demand in the 0 to 5 range, and only one person rated their mental demand above 65, in the 75-85 range.

Figure 64. Strip chart depicting NASA TLX mental demand scores from visit five (TLX.MD.V5) by treatment group.

From the strip chart (Figure 64) of the data, we see the gamified participants have final mental demand (TLX.MD.V5) scores that appear to average higher than the control participants’ final mental demand.
scores. Most notably the gamified group had four participants that rated their mental demand in excess of the maximum mental demand in the control group. We also see that the gamified group has a wider range of mental demand results compared to the control group.

The Welch’s t-test results with a p-value of 0.010. The resulting means for the control group and gamified group are, respectively, 14.4 and 41.0. The conclusion that can be drawn from these results is that there is indeed a difference between the two means, and specifically, the gamified group has a higher mental demand on average for the last visit.

This increased mental demand for gamified participants is promising for gamification as it works to keep each user engaged with the task, ideally lessening the potential for errors to occur due to distraction or boredom. Additionally, having a more interactive and mentally engaging task for assemblers can lead to better retention as this will help monotonous jobs feel less boring, so workers will be less likely to seek new employment opportunities.

5.5.2.2. Physical Demand at Visit 5

The physical demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” physical demand and 100 being “high” physical demand. Physical demand is measured by answering the question: “how much physical activity did you spend for this task?”

![Figure 65. Histogram of NASA TLX physical demand scores from visit five (TLX.PD.V5).](image)

The resulting histogram (Figure 65) of the values for physical demand at the last visit (TLX.PD.V5) shows two distinct clusters, one with values ranging from 0 to 25 and the other with values ranging from 35 to 65. The smaller cluster hosts majority of values, with the most common occurrence being the 0 to 5 range.
From the strip chart (Figure 66), we see that all occurrences of 0 physical demand lie within the control group. Additionally, the gamified group has three results that exceed the maximum value for the control group. Visually speaking, it appears as though the gamified group has a higher average physical demand score at the final visit than the control group.

Ultimately, the Welch’s t-test results with a p-value of 0.087. The resulting mean for the control group is 13.9, while the resulting mean for the gamified group is 29.0. Given these results, we conclude that gamification likely leads to increased physical demand for participants during the last visit.

As mentioned earlier in the results section for H3 (Physical Demand from Visit 1), it is possible that participants misperceived their mental and temporal demand for physical demand. It is peculiar to see that participants in the two groups rate their physical demand so differently, given that the physical demand of the task is exactly the same since the assembly tasks are exactly the same for the two groups. We believe that participants rate their physical demand higher in the gamified treatment group because they are trying to work faster and more efficiently to earn their badges. This feeling is more accurately related to temporal demand, but translates to perceived physical exertion.

5.5.2.3. Temporal Demand at Visit 5
The temporal demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” temporal demand and 100 being “high” temporal demand. Temporal demand is measured by answering the question: “how much time pressure did you feel in order to complete this task?”
Again, we see a varying degree of results, with temporal demand (TLX.TD.V5) scores ranging from 0 to nearly 85 (Figure 67). Similar to other NASA TLX scores at the last visit, the 0 to 5 range is the most common occurrence. From there, we see a mostly uniform distribution with approximately two values for other ranges. There are no participants that rated their temporal demand in the 45 to 55 range.

Figure 67. Histogram of NASA TLX temporal demand scores from visit five (TLX.TD.V5).

Figure 68. Strip chart depicting NASA TLX temporal demand scores from visit five (TLX.TD.V5) by treatment group.
Similar to the previous NASA TLX scores at the last visit, we see multiple gamified participants ratings exceeding the maximum rating for the control group (Figure 68). In this case, there are five participants. Additionally, we see that two control participants rate their temporal demand at 0, while no gamified participants rate their temporal demand at 0. The results for the gamified group are across a much larger range (from 5 to 85) compared to the control group (from 0 to 45).

The Welch's t-test results with a p-value of 0.010. The resulting means are 15.6 and 47.5 for the control group and gamified group, respectively. Our conclusion here is that gamification causes increased temporal demand (time pressure) ratings at last visit for the monotonous Lego assembly task used in this experiment.

The increased temporal demand for gamified treatment group participants at first appears to be a negative aspect of gamification, but we believe this moderate amount of time pressure is a positive attribute of gamification. The average temporal demand rating is 47.5, indicating a moderate amount of time pressure experienced by the gamified group participants. We believe that this amount of time pressure encourages participants to work more efficiently and to improve their cycle times without causing overwhelming stress and anxiety.

5.5.2.4. Performance at Visit 5
The performance is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “good” performance and 100 being “poor” performance. It is important to note the location of the endpoints with respect to their qualitative meanings. This is often the opposite of what would be expected. Performance is measured by answering the question: “how successful do you think you were in accomplishing the goals of the task?”

![Figure 69. Histogram of NASA TLX performance scores from visit five (TLX.P.V5).](image-url)
The resulting histogram (Figure 69) shows a vast majority of participants rating their performance (TLX.P.V5) in the 0 to 5 range. The maximum performance range falls in the 35 to 45 category. Aside from the 0 to 5 range, the performance ratings are primarily uniformly distributed with approximately two participants in each of the ranges.

From the strip chart (Figure 70), which separates the data into their respective treatment groups, we see numerous individuals in both groups rating their performance at 0. Again, we see three participants in the gamified group rating their performance scores higher than the maximum performance score of the control group. As a reminder, a lower performance rating indicates better performance, thus a performance rating of 0 indicates near perfect performance. Additionally, it is important to keep in mind that these performance ratings are self-rated by participants and are therefore, perceived performance ratings and are not dependent on their actual objective performance (cycle time or otherwise).

The resulting p-value from the Welch’s t-test for difference of means is 0.22 indicating no statistical difference between the two groups. The resulting means are 6.7 and 14.5 respectively. Here we cannot conclude that there is a difference between the means, and therefore it is possible that there is no impact on performance when gamification is applied compared to a control scenario.

Although there is no statistical difference between means, we can hypothesize a few relations based on the strip chart and mean results for the two groups. First, we see that the gamified participants rated their performance higher on average, indicating less confidence in their builds. We believe this occurred because participants in the gamified group had consistent feedback about their performance, allowing them to truly assess how well they performed. Whereas in the control group, these participants had very little indication about their performance, aside from when they completed the build with an error and were informed that they needed to return to building to resolve that error. The lower values for the control group are likely due to the notion: “ignorance is bliss.” This phrase refers to the idea that when
completing a task, when we do not know how well or how poorly we are doing, our minds tend to automatically assume we are doing well. On the flip side, consistent with the gamified group, when we have timely and consistent feedback, we often compare our results with our goals and aspirations. For some individuals, their goals may have been unattainable and therefore their perceived performance was rated high (poorly) because they had not reached the goal that they set for themselves. To provide a specific example supporting this statement, one participant stated that their goal was to complete the build faster than the researcher could disassemble the build. However, the researcher could disassemble the build in under three minutes, while the best build time across the entire study was in excess of ten minutes. Thus, the participant was setting unrealistic expectations for themselves.

This idea of poor perceived performance due to not reaching a goal is important for gamified scenarios that include goal setting. In our scenario, we set our badges to include easily attainable goals (completion of builds that participants were expected to complete) and flexible goals (the personal record badge). On the flip side, we could have set our badges to include goals that would have been extremely hard to attain, such as having a build time that breaks the ten-minute threshold. Implementing our GfM scenario with an extremely hard-to-achieve badge would be counterproductive in its attempt to motivate users. We recommend adopting a similar approach in industry. Carefully examining the badges to be included in an implementation of GfM is vital. At the same time, more research needs to be done to understand the direct and indirect implications of certain badges/goals on the targeted improvements. Likewise, implementers must iteratively reassess if the badges are working as intended. Finding the sweet spot of not-too-easy and not-too-hard badges or goals seems to be important for ensuring an engaging and motivating gamified scenario.

5.5.2.5. Effort at Visit 5
The effort is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” effort and 100 being “high” effort. Effort is measured by answering the question: “how hard did you have to work to accomplish your level of performance?”

Figure 71. Histogram of NASA TLX effort scores from visit five (TLX.E.V5).
The resulting histogram (Figure 71) is very sporadic and does not tell us much about the effort scores from visit five (TLX.E.V5). There are local valleys and peaks throughout the ranges, and the graph shows high variation in the self-rated effort scores for participants.

![Graph showing TLX effort scores from visit five by treatment group.](image)

Figure 72. Strip chart depicting NASA TLX effort scores from visit five (TLX.E.V5) by treatment group.

From the strip chart (Figure 72), we see that control participants have a lesser mean and lesser variance than the gamified group. The control group has a range from 0 to 35, while the gamified group has a range from 5 to 85. Similar to other strip charts for NASA TLX scores at visit five, we see six results for the gamified group that exceed the maximum value for the control group.

The Welch’s two sample t-test for difference of means results with a p-value of 0.011. The control group mean is 17.2 and the gamified group mean is 45.5. Based on the resulting p-value, we conclude that the gamified group has a statistically significant higher average effort score than the control group.

Although on first glance one would tend to believe that lower effort is desired to make a task easier, we conclude that higher effort is actually desired since it keeps users engaged with the task and leads to improved production efficiency. In this case of judging effort, it is important to recall the question asked to participants: “how hard did you have to work to accomplish your level of performance?” In our scenario with gamification, since these participants performed better than the control participants, we would directly expect that they exerted more effort. This is akin to grades in the classroom; generally students who receive an excellent grade exerted significantly more effort than students who received a poor grade. In many facets of life, the more effort put in, the greater the reward. Gamification shows itself to be a driver of increased effort and therefore improved performance.

5.5.2.6. Frustration at Visit 5

The frustration is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” frustration and 100 being “high” frustration. Frustration is measured by answering the question: “how insecure, discouraged, irritated, stressed, and annoyed were you during this task?”
The histogram (Figure 73) of frustration scores during the last visit (TLX.F.V5) best fits an inverted bell curve, with only a few values falling within the middle ranges (15 to 55) and additional values falling at the minimum and maximum ranges (0 to 15 and 55 to 65). Ultimately, the majority of participants rated their frustration in the 0 to 5 range. More than half of all participants rated their frustration as very low or moderately low with twelve participants rating their frustration below 25.

Figure 74. Strip chart depicting NASA TLX frustration scores from visit five (TLX.F.V5) by treatment group.
When looking at the strip chart (Figure 74), we see that the control group has a very strong cluster of results in the 0 to 10 range, with only one participant rating in excess of 10 (at frustration equal to 30). On the other hand, we see that the gamified group has a fairly uniform spread of results ranging from 0 to 65. The only repeated value is 65.

The Welch’s t-test results with a p-value of 0.0092. Respectively, the control group mean and gamified group mean are 7.2 and 34.0. From these results, we are extremely confident in stating that during the final visit, gamification has a significant impact on participants’ frustration, causing a higher level of frustration than that which the control group experienced.

We believe this increase in frustration for the gamified group might be explained by revisiting the idea of “last ditch efforts.” It was likely frustrating for participants in the gamified group to not achieve a better cycle time than they had hoped for by the completion of the study. On the other hand, participants in the control group were so accustomed to the task at hand by this point, that they were likely rather bored and underwhelmed by completing the build, therefore leading to very little or no sense of frustration.

With this being said, we believe that the moderate amount of frustration experienced by the gamified group of participants (average value of 34.0) is a healthy level of frustration that encouraged improvement and betterment throughout the duration of the study, especially at the last visit. This relates back to the final cycle times for each participant (CT15) and how the participants in the gamified group showed a sudden significant improvement from what we called “last ditch efforts” to achieve their personal record.

5.5.2.7. Conclusions on NASA TLX Scores at Visit 5

In conclusion, we obtained the following results for hypothesis 4 (H4), when looking at the TLX scores from the last visit (Table 20):

Table 20. Control group mean, gamified group mean, and resulting p-value of Welch’s t-test for variables related to NASA TLX scores at visit 5. P-values of significance are highlighted.

<table>
<thead>
<tr>
<th>Variable (score at visit 5)</th>
<th>Control group mean</th>
<th>Gamified group mean</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Demand</td>
<td>14.4</td>
<td>41.0</td>
<td>0.010</td>
</tr>
<tr>
<td>Physical Demand</td>
<td>13.9</td>
<td>29.0</td>
<td>0.087</td>
</tr>
<tr>
<td>Temporal Demand</td>
<td>15.6</td>
<td>47.5</td>
<td>0.010</td>
</tr>
<tr>
<td>Performance</td>
<td>6.7</td>
<td>14.5</td>
<td>0.22</td>
</tr>
<tr>
<td>Effort</td>
<td>17.2</td>
<td>45.5</td>
<td>0.011</td>
</tr>
<tr>
<td>Frustration</td>
<td>7.2</td>
<td>34.0</td>
<td>0.0092</td>
</tr>
</tbody>
</table>

From review and comparison of means for the two different groups for each of the NASA TLX score categories at the last visit, we see that all categories, except performance, return significant p-values indicating that gamification does have an impact on the perceptions of the participants for the given task. As discussed with each of the subsections, the increased ratings for participants in the gamified treatment group are in fact a positive attribute of gamification, as it allows the user to be fully engaged and mentally interactive with the otherwise boring and monotonous task at hand. This leads to improved quality and productivity for tasks of this nature (repetitive assembly tasks), ultimately improving a company’s bottom line.
5.5.3. Average NASA TLX Scores across all 5 Visits

In this subsection, we take a look at the average of each of the six NASA TLX scores from across all five visits. The average score across all five visits was found by summing each individual’s score for each visit, then dividing by five, just as an arithmetic mean would be calculated.

Many of the items worthy of discussion in this subsection are closely related to discussions covered in the previous subsection (5.5.2. NASA TLX Scores at Visit 5). As such, only results which are new and unique are discussed in this subsection.

5.5.3.1. Average Mental Demand across all 5 Visits

The mental demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” mental demand and 100 being “high” mental demand. Mental demand is measured by answering the question: “how much mental and perceptual activity did you spend for this task?” The average mental demand score across all five visits was found by summing each individual’s mental demand score for each visit, then dividing by five, just as an arithmetic mean would be calculated.

![Figure 75. Histogram of average NASA TLX mental demand scores (AvgMD).](image)

The resulting histogram (Figure 75) shows that the average mental demand score (AvgMD) for all participants is approximately normal, with a mode of 35-45. The minimum value is in the range of 0 to 75 and the maximum value is in the range of 65 to 75.
From the strip chart (Figure 76) of the data, we see that the control participants have more varying averages, ranging from approximately 0 to 65, while the gamification participants are closer together, with a range from approximately 30 to 67. Conversely to what we saw with the scores for the last visit, in this case we see that the control participants have multiple participants with mental demand scores lying below the minimum average for the gamification group participants, while having similar maximum values.

The Welch’s t-test results with a p-value of 0.043. The resulting means for the control group and gamified group are, respectively, 28.8 and 45.3. Thus, we conclude that the gamified group experienced statistically significant higher mental demand on average across all visits of the study.

5.5.3.2. Average Physical Demand across all 5 Visits
The physical demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” physical demand and 100 being “high” physical demand. Physical demand is measured by answering the question: “how much physical activity did you spend for this task?” The average physical demand score across all five visits was found by summing each individual’s physical demand score for each visit, then dividing by five, just as an arithmetic mean would be calculated.
Figure 77. Histogram of average NASA TLX physical demand scores (AvgPD).

The resulting histogram (Figure 77) of the values for average physical demand (AvgPD) show the range of 15 to 25 being the most common occurrence, with values ranging from as low as in the range of 0 to 5 and upwards of 65.

Figure 78. Strip chart depicting average NASA TLX physical demand scores (AvgPD) by treatment group.

From the strip chart (Figure 78), we see that the gamified participants again have multiple data points in excess of the maximum data point for the control group. We see many overlying data points between the
two groups but can generally see a difference in the two groups with the control group tending towards lower scores and the gamified group tending towards higher scores.

Ultimately, the Welch’s t-test results with a p-value of 0.095. The resulting mean for the control group is 17.9, while the resulting mean for the gamified group is 31.1. Given these results, we conclude that the difference between the means is significant and that the gamified treatment group experienced higher physical demand on average throughout the duration of the study.

5.5.3.3. Average Temporal Demand across all 5 Visits

The temporal demand is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” temporal demand and 100 being “high” temporal demand. Temporal demand is measured by answering the question: “how much time pressure did you feel in order to complete this task?” The average temporal demand score across all five visits was found by summing each individual’s temporal demand score for each visit, then dividing by five, just as an arithmetic mean would be calculated.

![Figure 79. Histogram of average NASA TLX temporal demand scores (AvgTD).](image)

The histogram (Figure 79) of the average temporal demand (AvgTD) across all five visits is unique. For values below 45, we see half of an approximately normal distribution, while values in excess of 45 appear mostly uniform. The average temporal demand scores range from the range of 0 to 5 and up to the range of 75 to 85.
Figure 80. Strip chart depicting average NASA TLX temporal demand scores (AvgTD) by treatment group.

Similar to the average physical demand scores (Figure 78), we again see that the gamified group has multiple data points in excess of the maximum control data point (Figure 80). Additionally, we can again see that the control data tends towards lower scores while the gamified data tends towards high scores.

The Welch’s t-test results with a p-value of 0.019. The resulting means are 30.4 and 54.5 for the control group and gamified group, respectively. Given our p-value is less than our desired alpha value of 0.10, we reject our null hypothesis and conclude that there is a difference between the two means. Considering the means for the two groups, we see that the gamified group has a statistically significant higher average temporal demand score across all visits of the study.

5.5.3.4. Average Performance across all 5 Visits

The performance is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “good” performance and 100 being “poor” performance. It is important to note the location of the endpoints with respect to their qualitative meanings. This is often the opposite of what would be expected. Performance is measured by answering the question: “how successful do you think you were in accomplishing the goals of the task?” The average performance score across all five visits was found by summing each individual’s performance score for each visit, then dividing by five, just as an arithmetic mean would be calculated.
The resulting histogram (Figure 81) shows that average performance scores (AvgP) are much more condensed than any other NASA TLX category, with results ranging only from the 0 to 5 range up to the 25 to 35 range. 13 participants resulted with an average performance score of less than 15, leaving six participants with an average performance score of more than 15, but less than 35.
From the strip chart (Figure 82), which separates the data into their respective treatment groups, we see two distinct clusters for the gamification group, one ranging from approximately 23 to 30 and the other ranging from approximately 0 to 10. On the other hand, the control group has a spread of data across values from approximately 2 to 23. The two separate clusters for the gamified treatment group makes it difficult to draw any conclusions from the data.

The resulting p-value from the Welch's t-test for difference of means is 0.25 indicating no statistical difference between the two groups. The resulting means are 9.2 and 14.8 respectively. The conclusion here is that there is no statistically significant difference between the two treatment groups.

Regarding the two clusters for the gamified treatment group, this may be an indicator for the phenomenon described in 5.5.2.4. Performance at Visit 5, where we discussed the idea that gamified participants may have thought they performed worse because they had not achieved their self-set goals for the study. The cluster in the gamified group with lower scores may be individuals who set reasonable or easy goals for themselves and felt accomplished through the study. While on the other hand, the cluster with higher average performance scores in the gamified group may be a set of individuals who set tougher goals for themselves typically and were then disappointed in their inability to achieve their self-set goals.

As mentioned previously, the performance ratings here are not clear indicators of how the participants actually performed, but these ratings are self-ratings provided by the participants by responding to the question: “how successful do you think you were in accomplishing the goals of the task?”

5.5.3.5. Average Effort across all 5 Visits
The effort is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” effort and 100 being “high” effort. Effort is measured by answering the question: “how hard did you have to work to accomplish your level of performance?” The average effort score across all five visits was found by summing each individual’s effort score for each visit, then dividing by five, just as an arithmetic mean would be calculated.

![Figure 83. Histogram of average NASA TLX effort scores (AvgE).](image)
The resulting histogram (Figure 83) shows a varying spread of average effort values (AvgE). Most interestingly, this is the first instance of a NASA TLX category not having any results in the 0 to 5 range. The average effort values range from 5 to 15 up to 65 to 75. The most common average effort scores are in the range of 25 to 35, closely followed by 35 to 45 and 55 to 65.

Figure 84. Strip chart depicting average NASA TLX effort scores (AvgE) by treatment group.

From the strip chart (Figure 84), we see that the control group average effort scores tend much lower than the gamified effort scores. Visually speaking, there is a clear difference between these two groups. Most notably, there are five gamified data point values that exceed the maximum control group value and four control group data point values that are less than the minimum gamified data point value.

The Welch’s two sample t-test for difference of means results with a p-value of 0.0048. The control group mean is 26.9 and the gamified group mean is 49.9. Based on the resulting p-value, we conclude that the gamified group indeed has a different mean than the control group, and that the gamified group tends to have much higher average effort scores throughout the duration of the study.

5.5.3.6. Average Frustration across all 5 Visits

The frustration is judged by participants on a scale from 0 to 100, in increments of 5, with 0 being “low” frustration and 100 being “high” frustration. Frustration is measured by answering the question: “how insecure, discouraged, irritated, stressed, and annoyed were you during this task?” The average frustration score across all five visits was found by summing each individual’s frustration score for each visit, then dividing by five, just as an arithmetic mean would be calculated.
The histogram (Figure 85) of the average frustration scores show a generally increasing number of occurrences from the 0 to 5 range up to the 15 to 25 range. Then, the remaining categories from 25 to 75 all have one occurrence, except the 45 to 55 range has three occurrences.

When looking at the strip chart (Figure 86) which breaks the two treatment groups apart, we see that the control group and gamified treatment groups have similar minimum values. However, the gamified group
has a larger count of values which exceed 30 than does the control group. We suspect from the visual to see a difference in the two means.

The Welch’s t-test results with a p-value of 0.087. Respectively, the control group mean and gamified group mean are 18.9 and 34.6. These results support our idea from the visual that there is a statistical difference between the means. Namely, the gamified group tends to experience more frustration on average than the control group. Most interestingly from this analysis, we observe that the mean for the gamified group is relatively moderate, at only 34.6.

5.5.3.7. Conclusions on NASA TLX Scores across all 5 Visits

In conclusion, we obtained the following results for H4, when looking at the average TLX scores across all five visits (Table 21):

Table 21. Control group mean, gamified group mean, and resulting p-value of Welch’s t-test for variables related to average NASA TLX scores across all five visits. P-values of significance are highlighted.

<table>
<thead>
<tr>
<th>Variable (average score)</th>
<th>Control group mean</th>
<th>Gamified group mean</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Demand</td>
<td>28.8</td>
<td>45.3</td>
<td>0.043</td>
</tr>
<tr>
<td>Physical Demand</td>
<td>17.9</td>
<td>31.1</td>
<td>0.095</td>
</tr>
<tr>
<td>Temporal Demand</td>
<td>30.4</td>
<td>54.5</td>
<td>0.019</td>
</tr>
<tr>
<td>Performance</td>
<td>9.2</td>
<td>14.8</td>
<td>0.25</td>
</tr>
<tr>
<td>Effort</td>
<td>26.9</td>
<td>49.9</td>
<td>0.0048</td>
</tr>
<tr>
<td>Frustration</td>
<td>18.9</td>
<td>34.6</td>
<td>0.087</td>
</tr>
</tbody>
</table>

For the remaining hypothesis, we will focus on using the average NASA TLX score across all five visits. The NASA TLX score difference from the first visit to the last visit is a representation of a participant’s improvement from start to finish of the study and the NASA TLX score at the last visit is a representation of a participant’s perceptions/feelings at the conclusion of the study. Therefore, the best representation of a participant’s perceptions/feelings throughout the duration of the study is the average NASA TLX score across all five visits.

5.6. Hypothesis 5

The fifth hypothesis (H5) of interest in this study centers on one other topic of interest to the study. This is formally stated as:

H5: A user’s personality has no effect on their perceptions/feelings.

To complete the analysis of this hypothesis, we analyzed a few different aspects of our data. First, we looked at i) the relations between a user’s MBTI attributes and their resulting average NASA TLX scores. Then, we compared ii) how the two competitive questions in our exit survey related to the average NASA TLX scores. Lastly, we analyzed iii) how an individual’s self-rated patience connected to their average NASA TLX scores.

5.6.1. MBTI and NASA TLX Scores

The MBTI assessment was used to identify personality traits of the participants. The MBTI results with a four-letter acronym. The first letter is always either an E or an I, relating to an extrovert or introvert personality type. The second letter is always either an S or an N, indicating sensing or knowing personality.
type. The third letter is always either a T or an F, indicating a thinking or feeling personality type. The fourth letter is always either a P or a J, indicating a perceiving or judging personality type.

For this analysis, we looked at each letter pair (E/I, S/N, P/J, and T/F) to see if they resulted in statistically significant different first visit NASA TLX scores (Score_V1) and different Average NASA TLX scores (AvgScore). As such, there were a total of 48 comparisons for this subsection (four letter pairs crossed with six first-time NASA TLX scores, plus four letter pairs crossed with six average NASA TLX scores). We have decided to only discuss the comparisons that resulted in significant p-values, with a conclusion subsection that lists the results for each analysis.

The first analysis of interest is the introvert/extrovert personality and physical demand scores for the first attempt.

![Figure 87. Strip chart depicting NASA TLX physical demand at visit one (TLX.PD.V1) by introvert/extrovert personality.](image)

From the strip chart (Figure 87), split by the two personality types, we see that the extroverts have a much smaller range of values than the introverts. The resulting p-value of the Welch’s two sample t-test is 0.044. The mean physical demand score for the extrovert personality type was 15.8 and the mean physical demand score for the introvert personality type was 33.8. Thus, we conclude that the introvert personality type rated their physical demand statistically higher than those in the extrovert personality type group.

Using the strip chart as visual support, although the t-test resulted with a significant p-value, it is difficult to state whether this is a clear conclusion to draw. Looking at the graph, we see that there is a distinct cluster of individuals in the introvert with physical demand visit one results in the range of 0 to 20, which is similar to the range of results for the extrovert personality. It is possible that the personality type does not actually have much to do with the way in which they rate their physical demand. A larger sample size here would be helpful in determining if this is a true phenomenon of statistically different means, or rather a coincidence in this case.
The next analysis of note is again with the introvert/extrovert personality, but in comparison to their average performance scores across all five visits.

![Strip chart depicting average performance scores (AvgP) by introvert/extrovert personality.](image)

Figure 88. Strip chart depicting average performance scores (AvgP) by introvert/extrovert personality.

Similar to the previous analysis, we see that extroverts tend to have a lower average performance rating, while introverts are again broken into two separate clusters, one comparable to the extrovert type and one much higher than the extrovert type (Figure 88). In this case, the resulting p-value is 0.074, with an average performance value for extroverts of 6.8 and an average performance value for introverts of 14.6. Again, a larger sample size or more uniform split (roughly ten in each group) may have provided a better representation of the true inclinations of the introvert vs. extrovert personalities in relation to the performance NASA TLX category scores.

The third analysis of interest is the first visit performance scores (TLX.P.V1) for the sensing and knowing personality types from the MBTI.
Figure 89. Strip chart depicting NASA TLX performance score at visit one (TLX.P.V1) by knowing/sensing personality.

From the strip chart (Figure 89), we see a wide range of values for the knowing group, while the sensing group is much closer together with many repeated values. Similarly with the chart, the Welch’s two sample t-test results with a p-value of 0.033. The mean first visit performance score for the knowing group is 33.8 and the mean first visit performance score for the sensing group is 8.2.

Individuals in the knowing group are characterized as individuals who are intuitive. Generally, they look at the big picture of a situation and use what they know to make inferences. On the other hand, “sensing” individuals tend to rely on their senses for information, and therefore take in information in an immediate, literal way. Therefore, in the case of the differing mean performance scores that we see here, we see one primary potential explanation. This being that those in the “knowing” personality group consider their performance throughout the first visit when responding to the question, including their views on all three builds. For nearly all participants, the first build was lengthy compared to subsequent builds since this was a new, and thus not yet a repetitive and monotonous, task for them. On the flip side, those in the sensing group were more likely rating their performance primarily on how they felt at the conclusion of the visit. For many participants, by the conclusion of the first visit, they had improved significantly from the first build to the third build since a learning curve is always steepest at the start.

Similarly, we see a significant p-value result for the difference in means for the knowing/sensing personality types in comparison to the average performance score across all five categories.
From the strip chart (Figure 90), the y-axis is on a much smaller scale than the previous comparison, thus the ranges of the two data samples are much more comparable. However, we still see a significant cluster in the sensing group with lower values than the knowing group. Ultimately, the Welch’s t-test results with a significant p-value of 0.025, and we conclude that the knowing personality type tends to have a higher (worse) mean (mean of all data points in a group) average (average of an individual’s performance across five visits) performance rating (18.3) than the sensing personality type (mean average performance rating = 7.7). The possible explanation here is similar to the explanation for the difference in means for the first visit. It is likely that the knowing group rated their performance worse on average than the sensing group because the knowing group looks at their overall performance, while the sensing group tends to only look at their most recent performance which tends to be better than the average due to learning and improvements over time.

5.6.1.1. Conclusions on MBTI and Average NASA TLX Scores

So far, we have discussed only the instances which resulted with significant p-values. In this subsection, we provide all analyses and their resulting p-values. The previously discussed significant p-values are highlighted in Table 22.

Table 22. Resulting p-value of Welch’s t-test for variables related to comparisons of MBTI and Average NASA TLX scores. P-values of significance are highlighted.

<table>
<thead>
<tr>
<th>MBTI Pair</th>
<th>NASA TLX Category</th>
<th>V1 or Avg</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introvert/Extrovert (I/E)</td>
<td>Mental Demand</td>
<td>V1</td>
<td>0.80</td>
</tr>
<tr>
<td>Introvert/Extrovert (I/E)</td>
<td>Mental Demand</td>
<td>Avg</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>Introvert/Extrovert (I/E)</strong></td>
<td><strong>Physical Demand</strong></td>
<td><strong>V1</strong></td>
<td><strong>0.044</strong></td>
</tr>
<tr>
<td>Introvert/Extrovert (I/E)</td>
<td>Physical Demand</td>
<td>Avg</td>
<td>0.22</td>
</tr>
<tr>
<td>Introvert/Extrovert (I/E)</td>
<td>Temporal Demand</td>
<td>V1</td>
<td>0.66</td>
</tr>
<tr>
<td>Introvert/Extrovert (I/E)</td>
<td>Temporal Demand</td>
<td>Avg</td>
<td>0.92</td>
</tr>
<tr>
<td>MBTI Type</td>
<td>Measure</td>
<td>Value 1</td>
<td>Avg</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>I/E</td>
<td>Performance</td>
<td>V1 0.54</td>
<td></td>
</tr>
<tr>
<td>I/E</td>
<td>Effort</td>
<td>V1 0.41</td>
<td></td>
</tr>
<tr>
<td>I/E</td>
<td>Frustration</td>
<td>V1 0.37</td>
<td></td>
</tr>
<tr>
<td>S/N</td>
<td>Mental Demand</td>
<td>V1 0.92</td>
<td></td>
</tr>
<tr>
<td>S/N</td>
<td>Physical Demand</td>
<td>V1 0.77</td>
<td></td>
</tr>
<tr>
<td>S/N</td>
<td>Temporal Demand</td>
<td>V1 0.74</td>
<td></td>
</tr>
<tr>
<td>S/N</td>
<td>Performance</td>
<td>V1 0.074</td>
<td></td>
</tr>
<tr>
<td>S/N</td>
<td>Avg</td>
<td></td>
<td>0.025</td>
</tr>
<tr>
<td>P/J</td>
<td>Mental Demand</td>
<td>V1 0.19</td>
<td></td>
</tr>
<tr>
<td>P/J</td>
<td>Physical Demand</td>
<td>V1 0.26</td>
<td></td>
</tr>
<tr>
<td>P/J</td>
<td>Temporal Demand</td>
<td>V1 0.65</td>
<td></td>
</tr>
<tr>
<td>P/J</td>
<td>Performance</td>
<td>V1 0.55</td>
<td></td>
</tr>
<tr>
<td>P/J</td>
<td>Effort</td>
<td>V1 0.72</td>
<td></td>
</tr>
<tr>
<td>T/F</td>
<td>Mental Demand</td>
<td>V1 0.83</td>
<td></td>
</tr>
<tr>
<td>T/F</td>
<td>Physical Demand</td>
<td>V1 0.35</td>
<td></td>
</tr>
<tr>
<td>T/F</td>
<td>Temporal Demand</td>
<td>V1 0.75</td>
<td></td>
</tr>
<tr>
<td>T/F</td>
<td>Performance</td>
<td>V1 0.34</td>
<td></td>
</tr>
<tr>
<td>T/F</td>
<td>Effort</td>
<td>V1 0.12</td>
<td></td>
</tr>
<tr>
<td>T/F</td>
<td>Frustration</td>
<td>V1 0.86</td>
<td></td>
</tr>
<tr>
<td>T/F</td>
<td>Avg</td>
<td></td>
<td>0.98</td>
</tr>
</tbody>
</table>

Considering the quantity of insignificant results, the MBTI does not appear to be a significant indicator of the perceptions/feelings of gamification. Therefore, when considering personality as being defined by the
MBTI categories, we conclude that we do not reject the null hypothesis for hypothesis 5 (H5). Thus, it is likely true that MBTI (as an indicator of personality) has no effect on the perceptions/feelings of gamification, both early in an experience and over an extended period of time.

5.6.2. Competitive Nature and Average NASA TLX Scores
For the second analysis relating to hypothesis 5 (H5), we used our exit survey questions to gauge participant’s personality. Specifically in this subsection, we focus on a participant’s self-rated competitiveness score. We had two questions built into the survey to assess competitiveness, one that focused on an individual’s competitiveness against themselves (you vs. you) and one that focused on an individual’s competitiveness against others (you vs. them). These were rated on a 5-point Likert scale.

We compared the competitiveness ratings to the average NASA TLX scores for each category, resulting in a total of twelve analyses (two competitive categories and six NASA TLX categories). To compare between the various attributes, we used a scatter plot, fit a linear model, and compared correlation values. In this subsection, we focus on the three comparisons that result in at least a mild correlation value (0.25 or higher and -0.25 or lower). For the other nine analyses, we include their correlation values in the conclusion section of this subsection.

One analysis of interest, the average temporal demand score vs. the self-rated competitive (you vs. you) rating, shows a clear positive relation between the two attributes.

Figure 91. Scatter plot of average temporal demand score (AvgTD) by competitiveness (you vs. you), with line to depict relation of the two variables.
From the graph (Figure 91), we see that participants who rated themselves highly competitive against themselves, also tended to feel increased temporal demand (time pressure). The resulting correlation value is 0.31.

Although we see a generally positive trend, we also see quite a few outliers which may be of significant importance to understanding the data. In particular, the outlier in the top right corner (competitive score of 5 and average TD score of approximately 85) is likely an important aspect of this trend line fitting this data. Without this outlier, the average for the competitive rating of 5 would be much lower, which would likely lead to the trend line being flatter and therefore less significant.

Aside from the potential impact of outliers, we can understand that an individual’s temporal demand is likely different based on their competitiveness, especially in considering their own expectations of themselves. This is an important consideration when developing gamification because it will be important to consider the competitiveness of the involved individuals, as well as the gamified scenario being implemented. If the individuals are already highly competitive, encouraging more competition may lead to significantly higher temporal demand, resulting in stress levels that are too high and therefore detrimental. On the other hand, for individuals who tend to be less competitive, inciting more competition may increase temporal demand, leading to that “sweet spot” of stress which encourages learning and improvements.

Another analysis of interest is the average self-rated performance score against the competitive you vs. you category.

![Figure 92. Scatter plot of average performance score (AvgP) by competitiveness (you vs. you), with line to depict relation of the two variables.](image)
From the graph (Figure 92), we see a few different things. First, we see the one outlier in the level 5 category for competitiveness. Additionally, we see that the level 4 category for competitiveness has two distinct clusters, one with relatively low scores and one with relatively high scores. The resulting correlation here is 0.29.

The graph and correlation value for average performance score and competitiveness (you vs. you) indicates that it is likely for participants with a higher competitiveness against themselves to believe their performance is worse (higher) than those with lower competitiveness scores. This is connected back to the previous conversations regarding individuals who tend to rate themselves with high (poor) performance likely being unsatisfied with their performance, as opposed to their performance being truly poor. Individuals who are competitive against themselves are more likely to set higher expectations and goals for themselves, and therefore, are more likely to view their performance as poor when they are unable to meet these expectations.

The final item of note is average frustration scores compared with competitiveness (you vs. you), and is also the most highly correlated comparison.

![Figure 93. Scatter plot of average frustration score (AvgF) by competitiveness (you vs. you), with line to depict relation of the two variables.](image)

In this case, we see a more obvious trend from just looking at the data points (Figure 93). Even without the fitted linear model, we could see and conclude a correlation between these two variables. The resulting correlation between these two variables is 0.38.
Here we conclude that for more self-competitive participants, higher average frustration is likely to occur. This seems logical because, for participants who are more competitive to improve upon their previous efforts, they will likely be more frustrated as they are unable to achieve their ambitious goals. Additionally, higher frustration aligns closely with having worse (higher) perceived performance. Generally, when someone thinks they have performed poorly, they are frustrated with themselves, especially when they are competitive with their previous self. Likewise, when someone is frustrated with their build due to mistakes, they may assume they performed poorly even if they performed well.

5.6.2.1. Conclusions on Competitive Nature and NASA TLX Scores

In the previous subsection we took a deep dive into the three most highly correlated analyses. Here we include a table (Table 23) of all correlation values for all analyses on competitive nature and average NASA TLX scores.

Table 23. Correlation values for competitive nature and NASA TLX scores. Correlation values of interest are highlighted.

<table>
<thead>
<tr>
<th>NASA TLX Category</th>
<th>Competitive Category</th>
<th>Correlation Value (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg Mental Demand</td>
<td>You vs. You</td>
<td>0.17</td>
</tr>
<tr>
<td>Avg Mental Demand</td>
<td>You vs. Them</td>
<td>-0.014</td>
</tr>
<tr>
<td>Avg Physical Demand</td>
<td>You vs. You</td>
<td>0.069</td>
</tr>
<tr>
<td>Avg Physical Demand</td>
<td>You vs. Them</td>
<td>-0.042</td>
</tr>
<tr>
<td>Avg Temporal Demand</td>
<td>You vs. You</td>
<td><strong>0.31</strong></td>
</tr>
<tr>
<td>Avg Temporal Demand</td>
<td>You vs. Them</td>
<td>-0.030</td>
</tr>
<tr>
<td>Avg Performance</td>
<td>You vs. You</td>
<td><strong>0.29</strong></td>
</tr>
<tr>
<td>Avg Performance</td>
<td>You vs. Them</td>
<td>0.14</td>
</tr>
<tr>
<td>Avg Effort</td>
<td>You vs. You</td>
<td>0.16</td>
</tr>
<tr>
<td>Avg Effort</td>
<td>You vs. Them</td>
<td>0.030</td>
</tr>
<tr>
<td>Avg Frustration</td>
<td>You vs. You</td>
<td><strong>0.38</strong></td>
</tr>
<tr>
<td>Avg Frustration</td>
<td>You vs. Them</td>
<td>0.078</td>
</tr>
</tbody>
</table>

From the table we see that exactly half (six of twelve) of the pairings have a correlation of approximately 0 (less than 0.1). Interestingly, two of the pairings with a correlation of approximately 0 are average temporal demand with competitiveness (you vs. them) and average frustration with competitiveness (you vs. them). So, even though temporal demand and frustration were impacted significantly by self-competitiveness, a self-rated value of competitiveness against others had seemingly no impact on their temporal demand and frustration for our scenario. Another interesting finding is that all instances of note here were for self-competitiveness. This is likely due to the test-case scenario being a highly individual situation with participants only competing against themselves. Therefore, individuals who are more competitive with themselves are likely to rate the various NASA TLX categories higher throughout the study than their counterpart since this scenario was built to interact with their self-competitive nature. Likewise, since this scenario did not have any sense of competing against others, this explains why there is no significant correlations between you vs. them competitiveness ratings and the various NASA TLX categories. Ultimately, this indicates that implementors of GfM with monotonous assembly tasks need to consider the two types of competitiveness of the future users so that they can ensure the type of gamification (self-competitive or competitive with others) is suitable for users. Again, we remind the
readers of the idea that there is a sweet spot for levels of stress to ensure users are actively engaged with the task at hand but not overwhelmed or discouraged.

5.6.3. Patience Rating and Average NASA TLX Scores
For the third and final analysis relating to hypothesis 5 (H5), we again utilized the results of a question from our exit survey to gauge the participant’s personality, specifically their patience. In the exit survey, participants were asked, “how patient are you with hard tasks?” They were provided with a 5-point Likert scale to respond to, ranging from “not at all” to “a great deal.” For our analysis, “not at all” corresponds to a patience rating of 1, and “a great deal” corresponds to a patience rating of 5.

We compared the patience ratings to the average NASA TLX scores for each category, resulting in a total of six analyses (one patience category and six NASA TLX categories). To compare between the various attributes, we used a scatter plot, fit a linear model, and compared correlation values. In this subsection, we focus on the two comparisons that result in at least a mild correlation value (0.25 or higher and -0.25 or lower). For the other four analyses, we include their correlation values in the conclusion section of this subsection.

The first analysis with a correlation of interest is average physical demand scores and self-rated patience.

![Figure 94. Scatter plot of average physical demand score (AvgPD) by patience, with line to depict relation of the two variables.](image)

From the graph (Figure 94), we see a significantly negative trend line, indicating that for those with higher patience levels, physical demand scores tended to be lower than for those with less patience. This is supported by the resulting correlation value of -0.36.
Considering our low sample size, this correlation value is fairly high and may indicate that an individual’s self-rated patience and ability to tolerate difficult tasks has an impact on their perceived physical efforts for completing a task. This is logical because for those with high tolerance for difficult tasks, they would likely be able to continue working through difficult points without feeling overexerted or as if they are spending additional physical efforts to complete the task at hand. Conversely, those with lower patience, are more likely to feel overwhelmed with tasks that require additional time and problem solving to complete, since their tolerance for completing hard tasks is lower. What this means for gamification is that, since we know gamification tends to increase average physical demand scores (5.5.3.2. Average Physical Demand across all 5 Visits), when implementing and designing GfM, we must also consider our target users’ patience and tolerance for hard tasks. Adding competitive layers and additional considerations to a task (such as game elements) for someone with low tolerance for difficult tasks should be considered carefully to ensure the user does not become overwhelmed and fatigued. In some instances though, gamification may help alleviate some fatigue and stress for users as it may act as a way to guide users through the process with frequent feedback, as well as to provide a slight distraction from the difficult task by reorienting the thinking towards achieving a goal or “winning” as opposed to simply completing the task.

The other analysis of interest is the comparison of self-rated patience scores and average frustration scores.

![Figure 95. Scatter plot of average frustration score (AvgF) by patience, with line to depict relation of the two variables.](image-url)
Again, we see a fairly negative relationship between patience and average frustration scores (Figure 95). This is accompanied with a correlation value of -0.3, indicating that for those with lower patience, they tend to experience higher levels of frustration, while those with higher patience, tend to be much less frustrated, on average.

This is essentially the definition of patience. It makes a lot of sense that for those who are less tolerant for difficult tasks, they would become more frustrated than those who are more tolerant. This relates back to GfM in a similar fashion as the previous discussion regarding average physical demand and patience. Ultimately, for an individual who is more patient, adding gamification to their monotonous assembly task may improve their attention to the task, especially in instances where challenging and stimulating goals are incorporated. Since these highly tolerant individuals are less likely to become frustrated by the new game elements of their job, the risks are low and would likely improve the experience for the individual, as well as productivity. On the other hand, for individuals with low self-rated patience, gamification should be considered carefully to ensure that these individuals do not become overwhelmed by the task. This may be an instance where implementing simpler, more easily-attainable goals would allow for the supportive nature of gamification without adding too much frustration and stress.

### 5.6.3.1. Conclusions on Patience Rating and Average NASA TLX Scores

Now that we have covered the two most significant analyses in this category, we will briefly discuss the other analyses and their correlations (Table 24).

Table 24. Correlation values for patience and NASA TLX scores. Correlation values of interest are highlighted.

<table>
<thead>
<tr>
<th>NASA TLX Category</th>
<th>Patience Category</th>
<th>Correlation Value (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg Mental Demand</td>
<td>Self-Rated Patience</td>
<td>-0.12</td>
</tr>
<tr>
<td>Avg Physical Demand</td>
<td>Self-Rated Patience</td>
<td>-0.36</td>
</tr>
<tr>
<td>Avg Temporal Demand</td>
<td>Self-Rated Patience</td>
<td>-0.11</td>
</tr>
<tr>
<td>Avg Performance</td>
<td>Self-Rated Patience</td>
<td>-0.24</td>
</tr>
<tr>
<td>Avg Effort</td>
<td>Self-Rated Patience</td>
<td>-0.24</td>
</tr>
<tr>
<td>Avg Frustration</td>
<td>Self-Rated Patience</td>
<td>-0.39</td>
</tr>
</tbody>
</table>

From the table, we see that there are no comparisons with a correlation of approximately 0, which is different from the previous subsection that had almost half of the analyses with a correlation at approximately 0. In all instances of NASA TLX vs. patience, we see a negative correlation, showing that lower patience scores indicate higher NASA TLX values, and higher patience scores results in lower NASA TLX values. As discussed with the two in-depth analyses of average physical demand and average frustration compared to self-rated patience scores, since an individual’s level of tolerance for difficult tasks tends to indicate their perceived load of the task, especially physical demand and frustration, it is important to consider the patience levels of potential users prior to designing the gamification.

### 5.7. Other Results

To provide a well-rounded analysis of all data collected as part of the experiment, we have included this subsection which looks at five other sets of information. Although these analyses do not directly relate to the hypotheses of the dissertation, these analyses provide context and support for understanding the full picture of our experiment.
First, we investigate how performance, meaning first attempt cycle time (CT1), difference from first to last cycle time (Dif_CT1.CT15), and last attempt cycle time (CT15) is affected by a user’s personality, specifically their MBTI personality measure. Second, we look at demographics, such as experience, role, or age, of our participants to see if they had an effect on a user’s performance. Third, we look at how responses vary for control or gamified participants to two questions from our exit survey, which asked participants about the amount of pressure they felt during the study to perform faster (time pressure) and to perform more accurately (accuracy pressure). Fourth, we were interested in the participant’s perception of gamification, so in our exit survey, we revealed the other treatment group (the one they were not a part of) to each participant, and asked them, “If you were part of the other group, how do you think you would have performed?” In this subsection, we discuss the results of this hypothetical question. Lastly, we provide a table of descriptive statistics for every variable included in our study.

5.7.1. Performance and Personality (MBTI)
Although we were initially only interested in how personality may impact an individual’s perceptions or feelings of gamification, we also wanted to see if there was any impact of personality on an individual’s performance. For this analysis, we looked at each pairing of the Myers Briggs Indicators (I/E, S/N, T/F, and P/J) with the first cycle time, difference between first and last cycle time, and the last cycle time. In total, there were twelve analyses.

Ultimately, no results were significant with a p-value less than or equal to 0.10. The table (Table 25) showcases the p-values for each of the tests.

Table 25. Resulting p-value of Welch’s t-test for variables related to comparisons of productivity and MBTI. P-values of significance are highlighted.

<table>
<thead>
<tr>
<th>Productivity Measurement</th>
<th>MBTI Pairing</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT1</td>
<td>I/E</td>
<td>0.30</td>
</tr>
<tr>
<td>Difference of CT1 and CT15</td>
<td>I/E</td>
<td>0.28</td>
</tr>
<tr>
<td>CT15</td>
<td>I/E</td>
<td>0.94</td>
</tr>
<tr>
<td>CT1</td>
<td>S/N</td>
<td>0.29</td>
</tr>
<tr>
<td>Difference of CT1 and CT15</td>
<td>S/N</td>
<td>0.28</td>
</tr>
<tr>
<td>CT15</td>
<td>S/N</td>
<td>0.82</td>
</tr>
<tr>
<td>CT1</td>
<td>T/F</td>
<td>0.37</td>
</tr>
<tr>
<td>Difference of CT1 and CT15</td>
<td>T/F</td>
<td>0.36</td>
</tr>
<tr>
<td>CT15</td>
<td>T/F</td>
<td>0.82</td>
</tr>
<tr>
<td>CT1</td>
<td>P/J</td>
<td>0.17</td>
</tr>
<tr>
<td>Difference of CT1 and CT15</td>
<td>P/J</td>
<td>0.18</td>
</tr>
<tr>
<td>CT15</td>
<td>P/J</td>
<td>0.56</td>
</tr>
</tbody>
</table>

With all analyses resulting in insignificant p-values, we conclude that an individual’s MBTI personality does not impact their performance ability for the monotonous assembly task utilized in this experiment.

5.7.2. Performance and Demographics
We were also interested in knowing if our participants’ demographics, such as student status (role), experience with Legos, or their age had an effect on their performance (cycle time data).
When looking at the impact of student status, we had two groups, full-time undergraduate students and full-time graduate students. We only had four graduate students, compared to 15 undergraduate students, so drawing conclusions here was difficult. While the graduate students tended to have a lower (better) first (Figure 96) and last cycle time (Figure 98) and a lower (worse) difference in cycle time from first to last build (Figure 97), it is difficult to know if this is the norm for graduate students to be better, or if we just happened to have graduate students in the study who were better than some undergraduate students. This is primarily of question because the best graduate students were not better than the best undergraduate students but fall within the range of the undergraduate results.

Figure 96. Strip chart depicting cycle time 1 (CT1) in seconds by student status.

Figure 97. Strip chart depicting difference from cycle time 1 to cycle time 15 (Dif_CT1.CT15) in seconds by student status.
Next, when looking at self-rated experience with Legos, participants responded with either average, good, or very good experience levels. Note that this self-rated experience was evaluated prior to the beginning of the study and was part of the survey that participants filled out to indicate interest in the study. Experience was not taken into account when selecting participants. Across the various experience ratings, we see varying results, indicating that previous experience with the task had little impact on the participants’ performance. Most interestingly, the average and very good categories resemble each other across all three performance categories, even though these are perceived to be very different experience levels (Figure 99, Figure 100, Figure 101).
Finally, we looked at the various cycle time data points and compared them with participant ages. We thought that age may have an impact on a participants’ prior knowledge of Legos, as well as a participants’ ability to learn Legos. We had suspected that those who had more recently assembled Legos (assumed to be younger) would have lower cycle times on average since they were likely to be more familiar with Legos. Ultimately, we see no trends of significance, and cycle time values are splattered across each graph (Figure 102, Figure 103, Figure 104).
Figure 102. Strip chart depicting cycle time 1 (CT1) in seconds by age.

Figure 103. Strip chart depicting difference from cycle time 1 to cycle time 15 (Dif_CT1.CT15) in seconds by age.
5.7.3. Time/Accuracy Pressure and Treatment Group

In addition to asking the participants to respond to temporal demand through the NASA TLX assessment each visit, at the conclusion of the study, as part of the exit survey, we asked participants to respond to two similar questions. The first question related to accuracy pressure, which asked “How much pressure did you feel during the experiment to complete the build correctly?” and allowed participants to respond on a 5-point Likert scale from “none at all” to “a great deal.” The second question related to time pressure, which asked “How much pressure did you feel during the experiment to complete the build faster each time?” and allowed participants to respond on a 5-point Likert scale from “none at all” to “a great deal.” In both cases, “none at all” corresponds to a value of 1 and “a great deal” corresponds to a value of 5.

We compared the responses of these two questions to the two different treatment groups. We suspected a few potential outcomes. For accuracy pressure, we thought it may be possible for control participants to feel more accuracy pressure than gamification participants since they had no feedback or guidance to support their build. On the other hand, we suspected that gamification participants may have felt more accuracy pressure since they were consistently being “judged” by the screen on whether or not the build was completed accurately.

Ultimately, for accuracy pressure, we see no difference between the two groups (Figure 105). The resulting mean for the control group was 2.8 and the resulting mean for the gamified treatment group was 3.2. The resulting p-value of the Welch’s t-test was 0.38.
Regarding time pressure, we had suspected it to be very likely that participants in the gamified group would experience higher time pressure to complete the build faster each time since they had the ability to earn their personal record badge each build, while participants in the control group had no idea of their cycle time at any point in the study.

Ultimately, we see a resulting mean for the control group of 3.1 and a resulting mean for the gamified group of 3.8. The Welch’s t-test returned a p-value of 0.14 which is greater than 0.10, and therefore insignificant. This is supported by Figure 106. Thus, there is no difference in the means between the treatment groups for time pressure.
What this means for gamification is that things aren’t always as they seem. GfM tends to have many premonitions primarily based on skepticism, with limited hard facts to support these ideas. Although the research for GfM is growing, we desperately need more empirical studies such as this to break down these barriers which prevent industry from implementing gamification for monotonous assembly tasks.

5.7.4. Other Group Performance
During the course of the experiment, one idea that continued to surface within the research team centers around the idea that we, as a society, just think gamification would be better. Empirical research prior to this study was limited, and frankly still is, so the ability to definitively state whether gamification is truly a positive and good experience for users was in question. So, we were curious what our participants, individuals who had experienced two very similar yet different scenarios, thought about gamification. At the end of the study, as part of exit survey, the very last interaction we had with participants was to explain what the “other” group experienced. Participants were informed that participants in the study had been split into two groups, a control group and a treatment group. It is important to note that the phrase “gamified” or “gamification” was never mentioned to participants in order to limit bias from the idea of “gaming” and fun. For participants in the control group, we showed a sample of the gamification interface and shared that participants in the “other group” (gamified) also had access to a stopwatch/timer throughout the duration of each build. For participants in the gamified group, we verbally explained that participants in the “other group” (control) had no interface, no progress bar, no stopwatch/timer, and no badges to earn. Participants were then asked to respond to the question, “If you were part of the other group, how do you think you would have performed?” on a 5-point Likert scale from “Much worse” to “Much better.” A response of “much worse” corresponded to a value of 1 and a response of “much better” corresponded to a value of 5.

To compare the results of the two groups, we created a strip chart graph and conducted a Welch’s two sample t-test.

![Figure 107. Strip chart depicting expected performance in the “other” group by treatment group.](image-url)
From the strip chart (Figure 107), we can see a fairly obvious trend that participants in the gamified treatment group tended to rate their expected performance much lower if they had been part of the control group, while participants in the control group rated their expected performance much higher if they had been part of the gamified group. Additionally, we see that the mean score for the control group was 3.8, while the mean for the gamified group was 2.2. The resulting p-value from the Welch’s t-test for difference in means was a staggering 0.00060. It is also interesting to note that exactly one participant in each group voted opposite from the other group members. For instance, in the control group, the average response is that participants believed they would have performed better in the other group (roughly a score of 4), while one participant indicated they would have performed worse (a score of 2). Similarly, in the gamified group, the average response was participants expecting to have performed worse in the other group (roughly a score of 2), while one participant indicated they would have performed better (a score of 4).

These results support our thoughts that, in general, people expect gamification to provide an improvement upon the control setting. One limitation of our research here is that we are unable to say if all aspects of our gamified scenario prompted these thoughts or which individual or combination of game elements would have been most impactful. For instance, it may have been the idea of having a stopwatch or not having a stopwatch that provided participants with their differing points of view. On the other hand, it may have been the entire mix of game elements (progress bar, feedback, stopwatch, and badges) that encouraged the participants to prefer the gamified experience over the alternative. Regardless, it is clear that, in general, people think gamification is the answer to improving a monotonous assembly task.

5.7.5. Descriptive Statistics for All Variables
To start, we look at multiple tables (Table 26, Table 27, Table 28, Table 29, Table 30, Table 31, Table 32) that depict how participants were split between various groups.

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>No. of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>9</td>
</tr>
<tr>
<td>Gamified</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 27. Count of participants in each group of the Introvert/Extrovert MBTI personality pairing.

<table>
<thead>
<tr>
<th>Introvert/Extrovert</th>
<th>No. of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introvert</td>
<td>6</td>
</tr>
<tr>
<td>Extrovert</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 28. Count of participants in each group of the Knowing/Sensing MBTI personality pairing.

<table>
<thead>
<tr>
<th>Knowing/Sensing</th>
<th>No. of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowing</td>
<td>8</td>
</tr>
<tr>
<td>Sensing</td>
<td>11</td>
</tr>
</tbody>
</table>
Table 29. Count of participants in each group of the Thinking/Feeling MBTI personality pairing.

<table>
<thead>
<tr>
<th>Thinking/Feeling</th>
<th>No. of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thinking</td>
<td>5</td>
</tr>
<tr>
<td>Feeling</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 30. Count of participants in each group of the Judging/Perceiving MBTI personality pairing.

<table>
<thead>
<tr>
<th>Judging/Perceiving</th>
<th>No. of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judging</td>
<td>11</td>
</tr>
<tr>
<td>Perceiving</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 31. Count of participants in each group of student status.

<table>
<thead>
<tr>
<th>Student Status</th>
<th>No. of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate, full-time</td>
<td>15</td>
</tr>
<tr>
<td>Graduate, full-time</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 32. Count of participants in each group of self-rated experience level.

<table>
<thead>
<tr>
<th>Experience with Legos</th>
<th>No. of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>6</td>
</tr>
<tr>
<td>Good</td>
<td>8</td>
</tr>
<tr>
<td>Very good</td>
<td>5</td>
</tr>
</tbody>
</table>

Next, we include a table (Table 33) for the mean, standard deviation, minimum, and maximum values for each variable in the data set.

Table 33. Descriptive statistics for every variable included in the study.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle Time 1</td>
<td>2431.34</td>
<td>913.1</td>
<td>1255</td>
<td>4500</td>
</tr>
<tr>
<td>Cycle Time 2</td>
<td>1799</td>
<td>557.1</td>
<td>928</td>
<td>3036</td>
</tr>
<tr>
<td>Cycle Time 3</td>
<td>1548.2</td>
<td>487.0</td>
<td>799</td>
<td>2576</td>
</tr>
<tr>
<td>Cycle Time 4</td>
<td>1500.7</td>
<td>373.9</td>
<td>958</td>
<td>2481</td>
</tr>
<tr>
<td>Cycle Time 5</td>
<td>1256.9</td>
<td>252.4</td>
<td>753</td>
<td>1669</td>
</tr>
<tr>
<td>Cycle Time 6</td>
<td>1131.8</td>
<td>234.3</td>
<td>701</td>
<td>1592</td>
</tr>
<tr>
<td>Cycle Time 7</td>
<td>1209.9</td>
<td>241.1</td>
<td>803</td>
<td>1782</td>
</tr>
<tr>
<td>Cycle Time 8</td>
<td>1100.2</td>
<td>296.0</td>
<td>677</td>
<td>1985</td>
</tr>
<tr>
<td>Cycle Time 9</td>
<td>1001.9</td>
<td>177.1</td>
<td>692</td>
<td>1356</td>
</tr>
<tr>
<td>Cycle Time 10</td>
<td>1082.8</td>
<td>193.6</td>
<td>743</td>
<td>1669</td>
</tr>
<tr>
<td>Cycle Time 11</td>
<td>944.4</td>
<td>165.5</td>
<td>721</td>
<td>1434</td>
</tr>
<tr>
<td>Cycle Time 12</td>
<td>933.1</td>
<td>171.6</td>
<td>699</td>
<td>1365</td>
</tr>
<tr>
<td>Cycle Time 13</td>
<td>999.8</td>
<td>142.8</td>
<td>777</td>
<td>1347</td>
</tr>
<tr>
<td>Cycle Time 14</td>
<td>883.3</td>
<td>122.4</td>
<td>745</td>
<td>1184</td>
</tr>
<tr>
<td>Cycle Time 15</td>
<td>845.7</td>
<td>131.4</td>
<td>663</td>
<td>1162</td>
</tr>
<tr>
<td>Difference (CT1 – CT15)</td>
<td>1585.7</td>
<td>872.2</td>
<td>560</td>
<td>3709</td>
</tr>
<tr>
<td>------------------------</td>
<td>--------</td>
<td>--------</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>Mental Demand Visit 1</td>
<td>55.0</td>
<td>22.6</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>Physical Demand Visit 1</td>
<td>28.2</td>
<td>22.6</td>
<td>5</td>
<td>65</td>
</tr>
<tr>
<td>Temporal Demand Visit 1</td>
<td>57.6</td>
<td>29.1</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td>Performance Visit 1</td>
<td>18.9</td>
<td>22.6</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>Effort Visit 1</td>
<td>49.2</td>
<td>17.7</td>
<td>25</td>
<td>80</td>
</tr>
<tr>
<td>Frustration Visit 1</td>
<td>39.7</td>
<td>31.3</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Mental Demand Visit 2</td>
<td>41.1</td>
<td>22.5</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>Physical Demand Visit 2</td>
<td>28.9</td>
<td>18.7</td>
<td>5</td>
<td>65</td>
</tr>
<tr>
<td>Temporal Demand Visit 2</td>
<td>45.5</td>
<td>23.3</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>Performance Visit 2</td>
<td>10.8</td>
<td>9.5</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Effort Visit 2</td>
<td>43.9</td>
<td>25.0</td>
<td>5</td>
<td>85</td>
</tr>
<tr>
<td>Frustration Visit 2</td>
<td>23.4</td>
<td>21.2</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Mental Demand Visit 3</td>
<td>31.2</td>
<td>22.6</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>Physical Demand Visit 3</td>
<td>23.4</td>
<td>15.6</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>Temporal Demand Visit 3</td>
<td>43.2</td>
<td>24.8</td>
<td>0</td>
<td>85</td>
</tr>
<tr>
<td>Performance Visit 3</td>
<td>10.8</td>
<td>13.6</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>Effort Visit 3</td>
<td>36.3</td>
<td>26.1</td>
<td>0</td>
<td>75</td>
</tr>
<tr>
<td>Frustration Visit 3</td>
<td>28.7</td>
<td>27.7</td>
<td>0</td>
<td>85</td>
</tr>
<tr>
<td>Mental Demand Visit 4</td>
<td>28.7</td>
<td>20.6</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td>Physical Demand Visit 4</td>
<td>21.8</td>
<td>21.6</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>Temporal Demand Visit 4</td>
<td>36.8</td>
<td>31.2</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>Performance Visit 4</td>
<td>9.5</td>
<td>11.4</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>Effort Visit 4</td>
<td>33.4</td>
<td>25.8</td>
<td>0</td>
<td>75</td>
</tr>
<tr>
<td>Frustration Visit 4</td>
<td>22.6</td>
<td>22.4</td>
<td>0</td>
<td>75</td>
</tr>
<tr>
<td>Mental Demand Visit 5</td>
<td>28.4</td>
<td>23.9</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>Physical Demand Visit 5</td>
<td>21.8</td>
<td>19.5</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Temporal Demand Visit 5</td>
<td>32.4</td>
<td>28.7</td>
<td>0</td>
<td>85</td>
</tr>
<tr>
<td>Performance Visit 5</td>
<td>10.8</td>
<td>13.8</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>Effort Visit 5</td>
<td>32.1</td>
<td>25.6</td>
<td>0</td>
<td>85</td>
</tr>
<tr>
<td>Frustration Visit 5</td>
<td>21.6</td>
<td>23.4</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td>Competitive (you vs you)</td>
<td>3.8</td>
<td>0.8</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Competitive (you vs them)</td>
<td>3.5</td>
<td>1.3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Accuracy Pressure</td>
<td>3.0</td>
<td>1.0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Time Pressure</td>
<td>3.4</td>
<td>1.0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Patience</td>
<td>3.7</td>
<td>1.0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Performance in Other</td>
<td>2.9</td>
<td>1.1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Age</td>
<td>20.8</td>
<td>3.8</td>
<td>18</td>
<td>32</td>
</tr>
</tbody>
</table>

In addition to these descriptive statistics, we also include line charts of the cycle time data in Appendix. The two line charts on the top half of the figure represent the nine control participants, with the first chart (top left) showcasing five participants, numbered 1-5, and the second chart (top right) showcasing the remaining four participants, numbered 1-4. Similarly, the two line charts on the bottom half of the figure represent the original ten gamified participants, with the first chart (bottom left) showcasing five participants, numbered 1-5, and the second chart (bottom right) showcasing the remaining five participants, numbered 1-5.
6. Holistic Discussion

In this section, we take a step back from each individual analysis and focus on the combination of the analyses to formulate general findings for each hypothesis. Then, we take a look at the combination of each hypothesis and their findings to provide a discussion on the general impacts of GfM on the mental state of workers and productivity from our empirical study which simulated a monotonous assembly task.

6.1. Individual Hypotheses

We begin our holistic discussion by focusing on each hypothesis individually to provide decisive conclusions on each topic. In the previous section, 5. Results and Discussions, we focused on the various analyses related to each hypothesis. In contrast, in this subsection, we look at the various associated analyses holistically to provide a comprehensive and generalizable summary of the findings for each hypothesis.

6.1.1. Hypothesis 1

The first hypothesis of interest in this study centers on the production performance of the participant at the start of the study. This is formally stated as:

H1: GfM has no effect on the cycle time the first time an assembly task is completed.

From comparing the first attempt cycle times (CT1) and the average first visit cycle times (AvgCT_V1) for each treatment group, we concluded that we cannot reject our hypothesis (H1). Therefore, we cannot state that there are any changes in cycle time caused by the implementation of GfM. Although we did not have statistically significant differences between the two group means, we did find that the mean for the gamified treatment group was lower than the control group. Additionally, the visualizations of the data showed that the gamified treatment group tended to have lower cycle times (see 5.2. Hypothesis 1).

Prior to our discussion on hypothesis 1 (H1) in the previous subsection, we highlighted assumptions that were made to support our analyses. Herein, we discussed whether the two samples had equal variance. By looking at the standard deviation values and strip charts (separated by treatment group) of two productivity measurements, one representing the “first time” variable (CT1) and the other representing the “over time” variable (Dif_CT1.CT15), we see that the control group has a much higher variance than the gamified group. This can also be seen in various strip charts throughout the previous section.

A larger sample size would have provided additional data, which would likely reduce the variance within each group. This may have allowed for a statistically significant result, or at least would have provided a truer representation of the data. It is known in statistics that larger sample sizes are more desirable (though often not cost efficient) because they provide a better representation of the population. Therefore, we recommend future, identical studies with additional participants to grow the sample size for this empirical study.

Ultimately, we conclude that, although our results do not show statistical significance, GfM seems to produce lower/improved cycle times the first time an assembly task is completed. Additionally, GfM seems to lessens/improves variance the first time an assembly task is completed.
6.1.2. Hypothesis 2
The second hypothesis of interest in this study centers on the production performance of the participant over a period of time. This is formally stated as:

H2: GfM has no effect on cycle time of an assembly task over time (5 days, 2-3 hours per day).

Here, the idea of “over time” corresponds to improvement from start to finish, as well as performance at the end of a span of time. We looked at four different analyses here, comparing the control group to the gamified group results for i) difference between first and last cycle time (Dif_CT1.CT15), ii) difference between first and last visit cycle time (Dif_AvgCTV1.AvgCTV5), iii) last attempt cycle time (CT15), and iv) average last visit cycle time (AvgCT_V5). The results showed statistical significance only for the difference in means for the last attempt cycle time (CT15). In this one instance, we can definitively state that GfM has an effect on cycle time of an assembly task for the last attempt (CT15). In the other three instances, although not statistically significant, we see that the gamified treatment group has a lower average result in all three individual analyses. For the analyses that focus on the improvement (difference in cycle times) over the study, a lower average result indicates that there was less improvement. However, this is likely attributed to the lower initial cycle times (CT1) as discussed in 5.2. Hypothesis 1, since the final cycle times (CT15) are still lower on average for the gamified treatment group.

Ultimately, we conclude that, although our results only show one instance (of four) of statistical significance (rejecting H2), GfM tends to produce lower/improved cycle times at the end of a series of assembly tasks. Additionally, GfM tends to produce less improvement over time, which is attributable to the lower cycle times (CT1) at the start of the study.

Future research related to this hypothesis should investigate improvement as a ratio in order to have more comparable values given the difference in cycle times from the beginning. It would also be interesting to investigate how knowing that an attempt is the “last” attempt may affect a participant’s performance. We discussed in the results section that we believe participants in the gamified treatment group may have had a last attempt cycle time (CT15) significantly lesser than the control group in a final effort to earn their personal record badge. If participants were not aware this was their last attempt, we wonder whether they would have shown similar motivation to achieve such a low final cycle time (CT15). In all instances of this study, participants knew that this (build 15) was their final attempt. The difference is that in the gamified group, they had the potential to earn a personal record badge, as well as had knowledge of their progress through a progress bar and stopwatch. Meanwhile in the control group, participants were not eligible to earn a personal record badge and had no concept of their cycle time.

6.1.3. Hypothesis 3
The third hypothesis (H3) of interest in this study centers on the psychological effects of gamification on the participant at the start of the study. This is formally stated as:

H3: GfM has no effect on the perceptions/feelings of users the first time an assembly task is completed.

For this analysis, we used the NASA TLX assessment categories (mental demand, physical demand, temporal demand, performance, effort, and frustration) to represent the perceptions/feelings of users. We used the NASA TLX assessment results from visit one to represent the “first time” aspect of the hypothesis.
Our results showed that only temporal demand showed statistically significant differences between the control group and the gamified group. However, in all cases, we saw that the mean values were higher for the gamified treatment group than for the control treatment group.

When designing the study and defining hypothesis 3 (H3), we suspected that the gamification-supporting outcome would be lower values for each category, indicating less perceived workload on the participants. However, after further consideration and analysis of the results, we believe that moderately higher values for each category for gamification aligns with the intent of gamifying a task and are thus in support of gamification. As previously discussed, in order to keep operators during repetitive assembly tasks engaged, there must be some aspect of risk and/or motivation. It is a well-known phenomenon in industry, that when (assembly) tasks are too monotonous, the risk of mistakes increases (Wollter Bergman et al., 2021). Furthermore, the boreout phenomenon, often associated with office tasks, applies to this as well and can lead to significant health issues or silent quitting of the employee (Hod, 2022). While a lot of the prior work on GfM focused on the physical impact of repetitive (assembly) tasks, the mental aspect is considered of similar importance in this analysis. Generally, in daily, monotonous assembly tasks, the risk associated with a mistake is in most cases simply to scrap the build or to disassemble and reassemble. This leads to a drop in productivity, can ultimately cost the company money, or damage their relationship with customers among other things. In some companies, especially ones with large throughputs of similar products, such as healthcare products or automotive manufacturing, if the defects are not caught prior to delivery to the customer, this may result in a significant safety issue for consumers. Although this concern falls back to the company, it generally stays at the upper management level, rarely making it back to directly impact any individual employee on the shopfloor. Therefore, by employing gamification, the new risk to assemblers becomes “losing the game.” It adds a sense of both urgency and need for high-quality assembly before even leaving the assembler’s station, which then leads to improved productivity and quality at the source, instead of further down the assembly line.

Ultimately, we conclude that GfM has a significant impact on temporal demand, which incites assemblers to work more quickly, in turn reducing cycle times and improving productivity. It is important to note that our cycle time data in this study incorporates ensuring perfect quality of a build, and therefore, low cycle times indicate faster building and/or improved quality while high cycle times indicate slower building and/or worse quality that required rework. Additionally, we conclude that GfM tends to increase perceived workload for assemblers across all categories of the NASA TLX assessment. We see this as a positive for GfM because it keeps assemblers actively engaged with the task at hand, which is especially important in the early stages of learning an assembly task as it aids in faster learning rates. In addition, we suspect that gamification reduces the risk of boreout and associated negative effects by adding a dynamic element to a monotonous task, thus keeping the operator mentally engaged.

6.1.4. Hypothesis 4
The fourth hypothesis (H4) of interest in this study centers on the psychological effects of gamification on the participant over a period of time. This is formally stated as:

H4: GfM has no effect on the perceptions/feelings of users for an assembly task over time.

Similar to hypothesis 3 (H3), perceptions/feelings were derived from the NASA TLX assessment score. Similar to hypothesis 2 (H2), we considered “over time” in multiple ways, including i) difference/improvement from start to end of the study (Score_V1-Score_V5), ii) score at the end of the
study (Score_V5), and iii) average score throughout the study (AvgScore). For this holistic discussion on H4, we focus on the average NASA TLX scores across all five visits of the study, as this provides the best measurement of the participants’ perceptions/feelings throughout the study, and therefore a more general insight of the impacts of GfM. From the analysis of the results, we find that five of the six NASA TLX categories returns a significant p-value, indicating that we would reject our hypothesis and conclude that GfM has an effect on the perceptions/feelings of users. The one NASA TLX category that does not show any significant difference in the means of the control group and the gamified group is performance, however we still see a higher mean performance score for the gamified group than the control group.

Ultimately, we find that GfM has a significant effect on the perceptions/feelings of users throughout the study. GfM certainly causes an increase in mental demand, physical demand, temporal demand, effort, and frustration, and tends to cause a perceived-to-be worse level of performance, despite the fact that cycle time data showed that participants with GfM performed better than those without GfM. These increases in load are all only mildly higher than the control setting, with all gamified treatment means less than 25 points higher than the control group means, on a scale of 100 points. We argue that the increases in the various types of psychological load are a positive aspect of gamification as it encourages users to stay more engaged and mentally active in the task at hand, leading to improved performance, without causing extreme and/or unnecessary levels of stress.

6.1.5. Hypothesis 5
The fifth hypothesis of interest in this study centers on one other topic of interest to the study. This is formally stated as:

H5: A user’s personality has no effect on their perceptions/feelings

For this analysis, we utilized the NASA TLX assessments to represent the perceptions/feelings, similar to hypotheses 3 (H3) and 4 (H4). For personality, we utilized the MBTI results, competitive nature survey questions, and patience rating survey questions.

The MBTI results included an analysis of 48 comparisons, with only four resulting in significant p-values. Therefore, with so many results to consider and very few significant results, we conclude that MBTI personality has no generalizable effect on the perceptions/feelings of participants.

When looking at competitive nature, we found three (of six) instances of self-competitive nature (Competitive_UvU) having a significant impact on the NASA TLX category. These categories were temporal demand, performance, and frustration. We found that the competitive nature of “you vs. them” (or you vs. others) had no significant effect on any NASA TLX category. Thus, we conclude that an individual’s self-competitive nature has an effect on their psychological state during an assembly task, which is likely to interfere with the effects of GfM.

The self-rated patience (Patience) rating for an individual was found to be negatively correlated to physical demand and frustration, indicating that as patience levels decreased (less patient), physical demand and frustration levels rose. With regards to gamification, this indicates that when designing GfM applications, it is important to take into consideration the level of tolerance that the potential user will have for more difficult scenarios. As discussed with hypothesis 4 (H4), inciting some level of challenge provides a helpful level of stress, but exceeding a certain threshold (which varies for each individual) can lead to detrimental outcomes including burnout. This may lead to secondary effects such as high employee turnover.
Based on the conclusions herein, we recommend that those interested in implementing GfM should first consider the personality of the potential users by understanding their self-competitive nature (Competitive_UvU) and self-rated level of patience (Patience) with hard tasks. Knowing these personality attributes of the potential users will inform the initial design parameters of the gamified scenario. Future research in GfM should focus on diving deeper into developing a framework similar to our classification of game elements framework (Keepers, M., Nesbit, I., Wuest, T., 2022) to support the design decisions based on a future user’s personality.

Ultimately, we conclude that an individual’s self-competitive nature and their self-rated level of patience has an effect on their perceptions/feelings, while an individual’s MBTI personality has no effect.

6.2. Everything Considered
In this subsection, we look at all individual analyses and final conclusions for each of the five hypotheses of this study to provide a final discussion by connecting the results of the empirical study back to the identified research gaps which motivated the study.

Across all hypotheses of interest, we were either able to definitively reject the hypothesis or had results that showed a trend towards rejecting the hypothesis. In cases where we are not able to definitively reject the hypothesis, but instead relied on what the data appeared to be showing us, we believe additional samples of the data could aid in understanding the true nature of GfM for assembly tasks.

To reflect, the research gap that this experimental setup aimed to address was the iii) desperate need for additional empirical data of GfM, on both the psychological and production effects. This research directly addressed this research gap by providing clear analysis of empirical primary research data and thorough discussions to support the understanding of the data. Hypotheses 1 (H1) and 2 (H2) provide empirical data on the production effects of GfM, while hypotheses 3 (H3), 4 (H4), and 5 (H5) provide empirical data on the psychological effects. In an indirect fashion, this study supports future generation of additional empirical data of GfM by providing a detailed, transparent, and structured methodology (replicable) that others across the globe can implement and expand upon. Thereby, allowing for future research teams to add to the data set without starting from scratch (they would already have 20 data points from this research). One of the arguably hardest parts of researching GfM is being able to measure the long-term effects of GfM. Here we are able to provide 15 build attempts that were spread over roughly a five-week time period. Ideally, given that the methodology is already clearly spelled out for how to run a similar study, others could immediately start implementing a similar empirical study without much additional planning and preparation, allowing for additional study time, resulting in longer-term studies on the effects of GfM.

This empirical study, in addition to contributing to research gap iii), also addresses the research gap iv) a need for the development of step-by-step guidelines for implementing GfM by providing concrete recommendations for design considerations. These are included throughout the dissertation, but specifically from the empirical study results, one example is the recommendation to consider the potential user’s self-competitiveness and self-rated patience levels prior to design, since these personality traits have an impact on the user’s psychological load while completing an assembly task.

This section focused solely on connecting the research gaps to the empirical study. The following section, 7. Conclusion, discusses additional research gaps which are addressed by the entirety of the dissertation.
7. Conclusion

Gamification has been widely implemented in fields such as healthcare, marketing, and education, but has seen little interest and research for its application in manufacturing to date. In this dissertation, the overall goal of the research was to explore and establish GfM as a valid industrial productivity enhancement technique, by focusing on closing the four primary research gaps which were identified through a comprehensive and critical literature review (Keepers et al., 2022). The four identified research gaps were, a need for:

i. The use and acceptance of the Deterding definition,
ii. A clearer definition for the alternative game element terms,
iii. Additional empirical future research, and
iv. The development of step-by-step guidelines for implementing GfM.

Building on these research gaps and our understanding of the current body of literature, we developed an empirical study focused on the following five hypotheses of interest:

H1: GfM has no effect on the cycle time the first time an assembly task is completed.
H2: GfM has no effect on cycle time of an assembly task over time.
H3: GfM has no effect on the perceptions/feelings of users the first time an assembly task is completed.
H4: GfM has no effect on the perceptions/feelings of users for an assembly task over time.
H5: A user’s personality has no effect on their perceptions/feelings.

The empirical study centered around a mock assembly task which had 20 participants assemble a Lego Telehandler (build kit #42133). To simulate the monotony of a typical production facility, participants assembled the same kit 15 times, spanning five separate visits. At each visit, we collected cycle time data for three builds and one NASA TLX assessment, which included six data points (one for each of the following: mental demand, physical demand, temporal demand, performance, effort, and frustration). The cycle time data represented the production efficiency by encompassing both productivity, i.e., how fast the participant could assemble, and quality, i.e., how accurately the participant could assemble. The NASA TLX assessment represented the perceptions/feelings of the users by measuring how intensely they felt for each category of the assessment during each visit. A user’s personality was measured two ways, i) through the Myers Briggs Personality Type Indicator and ii) through a series of questions in an exit survey that centered on competitiveness and patience.

Nearly 100 unique analyses were conducted to gather a comprehensive picture on the propensity of GfM for monotonous assembly tasks. With thorough analyses consisting of visualizations, t-tests, and other considerations, we provide an in-depth discussion on each hypothesis in 6. Holistic Discussion.
In summary, we made the following conclusions (Table 34):

Table 34. Condensed list of conclusions for each of the five primary hypotheses (H1-5) of the study.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>GfM has no effect on the cycle time the first time an assembly task is completed.</td>
</tr>
<tr>
<td></td>
<td>• No statistical significance.</td>
</tr>
<tr>
<td></td>
<td>• Tends to produce lower/improved cycle times the first time an assembly task is completed.</td>
</tr>
<tr>
<td></td>
<td>• Tends to produce lesser/improved variance in cycle times the first time an assembly task is completed.</td>
</tr>
<tr>
<td>H2</td>
<td>GfM has no effect on cycle time of an assembly task over time.</td>
</tr>
<tr>
<td></td>
<td>• Statistical significance for final attempt cycle time (CT15), no other statistical significance.</td>
</tr>
<tr>
<td></td>
<td>• Tends to produce lower/improved cycle times at the end of a series of assembly tasks.</td>
</tr>
<tr>
<td></td>
<td>• Tends to produce less improvement over time, which is attributable to the lower cycle times at the start of the study.</td>
</tr>
<tr>
<td>H3</td>
<td>GfM has no effect on the perceptions/feelings of users the first time an assembly task is completed.</td>
</tr>
<tr>
<td></td>
<td>• Statistical significance for temporal demand (TD), no other statistical significance.</td>
</tr>
<tr>
<td></td>
<td>• Tends to increase perceived workload across all NASA TLX categories, which is a positive attribute of GfM.</td>
</tr>
<tr>
<td>H4</td>
<td>GfM has no effect on the perceptions/feelings of users for an assembly task over time.</td>
</tr>
<tr>
<td></td>
<td>• Statistical significance for all NASA TLX categories, except performance.</td>
</tr>
<tr>
<td></td>
<td>• Tends to cause a perceived-to-be worse level of performance, but this is countered by the positive results from H1 and H2.</td>
</tr>
<tr>
<td>H5</td>
<td>A user’s personality has no effect on their perceptions/feelings.</td>
</tr>
<tr>
<td></td>
<td>• No statistical significance for MBTI and any NASA TLX category.</td>
</tr>
<tr>
<td></td>
<td>• Moderately positive correlation for self-competitive nature and temporal demand, performance, and frustration.</td>
</tr>
<tr>
<td></td>
<td>• Minimal to no correlation for self-competitive nature and other three NASA TLX categories, and minimal to no correlation for competitiveness against others with any NASA TLX category.</td>
</tr>
<tr>
<td></td>
<td>• Moderately negative correlation for self-rated patience for physical demand and frustration.</td>
</tr>
<tr>
<td></td>
<td>• Minimal to no correlation for self-rated patience and the other four NASA TLX categories.</td>
</tr>
</tbody>
</table>

Ultimately, we conclude that there is evidence that indicates GfM is worthy of further research. Although we were not able to provide a blanket statement of significant improvements in productivity or mental wellbeing for participants exposed to gamification in our study, we were able to observe that GfM, in many cases, tends to improve various aspects of interest. Therefore, we strongly encourage the continuation and expansion of research efforts for GfM to better understand the true effects and implications of utilizing GfM in different environments, use cases, and tools.

7.1. Contributions to the Body of Knowledge

This dissertation addresses all four of the identified research gaps, each through multiple avenues.

First, the use and acceptance of the Deterding definition is supported through the in-depth consideration of various definitions for gamification in 2. State of the Art. It is a definition which is clear enough to describe gamification while also simple enough to allow for the flexibility that gamification requires to be successful.
Second, a clearer definition for the alternative game element terms is addressed through the creation of the classification framework as described in 3.1.1. Game Elements and my peer-reviewed conference article (Keepers, M., Nesbit, I., Wuest, T., 2022). To support the validation of this framework, we utilized two separate approaches of validating with literature, and also utilized the framework for the identification of game elements for this empirical study. Given the conclusions of our study, that GfM is likely a positive tool for productivity improvement in industrial settings, we can conclude that the game elements chosen according to the classification framework were appropriate for our setting.

Third, through our detailed explanation of our methodology, we support additional empirical future research by modeling a specific setup for future empirical studies that could be utilized anywhere in the world. More directly, we contribute to the small but growing body of empirical research for GfM by providing primary research data collection and analysis with thorough discussions and conclusions in this dissertation.

Fourth, we support the development of step-by-step guidelines for implementing GfM by i) detailing the use of the classification framework and ii) providing real-world context to all explanations and discussions. We were not able to provide explicit step-by-step guidelines for the implementation of GfM given the overall immature state of research on GfM. We express the need for additional research to support this development in 7.2. Limitations and Future Work. However, our work does provide explicit recommendations for selecting game elements, which is an early stage of designing and implementing GfM through our classification framework. Our work also explicitly provides real-world context and considerations for the implementation of GfM throughout 5. Results and Discussions and 6. Holistic Discussion.

Through the work included in this dissertation, GfM has proven to be a potentially valid industrial productivity enhancement technique through selected statistically significant results and several additional positive results. With the continuation of similar research, GfM will be further supported and shaped by a wealth of knowledge that provides further insights regarding its validity as a technique for industrial productivity enhancement.

7.2. Limitations and Future Work
An important contribution of every research work is a critical reflection on the limitations of the work as well as an outlook of future research opportunities emerging from the lessons learned. In this study, the limitations are grouped into two broader categories, i) study design and ii) participants. We elaborate more on each of the limitations in the following paragraphs, but first provide an overview for how they fall into the broader categories. The first limitation focuses on the length of time required by participants and this falls into both categories of limitations. The second limitation best belongs to the limitation category of participants and focuses on the limitations that are inherent of human subject research. The third limitation focuses solely on the study design and discusses the limitations that are inherent to a lab-based study as opposed to an in-situ factory experiment.

This study required participants to attend five different sessions, each lasting anywhere between 30 minutes and over 2 hours, depending on the individual’s performance. This required time commitment made it difficult to obtain and retain participants in the study. Additionally, as this study involved human research participants, subjectivity and external factors were minimized as best as possible but may have not been completely eliminated from or accounted for in the data analysis. For example, one participant
repeatedly attempted to hold conversation with the researcher during assembly times. The researcher repeatedly requested the conversation to end, but the participant continued to partake in conversation. Objectively, it did not appear to impede on the participant’s results whether through observations at the time of collection or later during data analysis, so the participant was not removed from the data pool.

Another consideration for working with human research participants across multiple days is how each participant is feeling on any given day that they arrived to the study. From the data analysis, we did not see any clear signs of exceptionally good or exceptionally bad days, so none were removed from the study, but they may have existed. Another limitation of the study is the relatively small sample size. This is frequently a concern with studies of gamification, especially those that are interested in examining the effects of gamification over an extended period of time, as it is often difficult to retain participants for the continuation of the study. At the same time, this limitation is difficult to overcome given the limited resources generally available for a research study. Hence, the suggestion to augment the data set by collecting additional data points following the presented methodology (with the caveat of ensuring similar study conditions and biases).

The other key limitation of this study is that it was conducted with students in a lab environment, as opposed to working professionals in a factory setting. Although we could not work in situ for this experiment, we ensured a study that closely aligned with a typical manufacturing assembly task by providing a clear set of work instructions, little one-on-one supervision, and using tools the participants were familiar with. We purposefully did not include an assembly task that would require participants to use any tools to ensure that all participants had similar preparation for the tasks at hand. Although some participants may have had prior experience with the task (in this case assembling a Lego kit), we saw in 5.7.2. Performance and Demographics that this had no impact on their performance for the study.

Opportunities for future work stemming from the results of this dissertation are vast, as the basis of empirical studies for GfM is currently extremely limited. There are a number of promising routes to continue with this research. First, future work may consider focusing on adding to this current data set by running similar studies based on the detailed methodology included herein. By adding to this data set and then analyzing the conglomerate data, this aids in reducing the limitation of limited sample sizes with lesser time and effort requirements from the researchers, as well as any bias included by the local study setup and observers. Second, future work may consider formulating similar studies utilizing an in-situ assembly task with factory workers. Although a few studies have been conducted with questionnaires or surveys and factory workers, to the best of our knowledge, little to no empirical studies have been conducted with GfM assembly and factory workers (aside from Korn’s work with sheltered workplaces). More specifically, a study which showcases the implementation of GfM by identifying game elements utilizing the classification framework described in 4.1. Preparation and Development would be of value to the GfM research community. This dissertation describes the purposeful selection of game elements for the lab-based study, however an extended study of utilizing the classification framework for a real factory setting is still necessary.

Looking at the finer details of the methodology, future work should consider the different possible interpretations of the “perceptions/feelings” of workers by looking at other measures of perceived workload and how the users feel while utilizing GfM. One consideration may be measuring physiological data such as heart rate, perspiration, or eye movement to augment the richness of data. Similarly, additional considerations for identifying one’s personality would be of interest for future work, possibly
by using other personality indicators or additional survey questions. Future work should also consider
different interpretations of the “over time” aspect of the hypotheses included in this dissertation, by
running studies with more builds and a longer timeframe. This dissertation takes a first attempt at looking
into the long-term implementation of GfM for assembly, but a longer-term study may provide different
or additional results on the efficacy of GfM for assembly. The final piece of future work that plagues us is
understanding why many participants, especially in the gamified treatment group, produced such
significantly lower final attempt cycle times (CT15). While having a lower cycle time is desirable, we were
surprised during data analysis to see the drastic change in cycle times that occurred between the 14th and
15th build. This is discussed in further detail in 5.3.3. Final Attempt Cycle Time. With the data we collected,
we had not anticipated this phenomenon, and therefore, were unable to discuss this with participants or
collect data to understand why this occurred.
8. References


9. Appendix

Table 35. Count of papers, organized by overlapping industries and work processes within the reviewed papers.

<table>
<thead>
<tr>
<th>Industry</th>
<th>Work Process</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>air transport</td>
<td>general operations</td>
<td>1</td>
</tr>
<tr>
<td>automotive</td>
<td>bolt-tightening</td>
<td>2</td>
</tr>
<tr>
<td>automotive</td>
<td>general assembly</td>
<td>2</td>
</tr>
<tr>
<td>electronics</td>
<td>general manufacturing</td>
<td>2</td>
</tr>
<tr>
<td>equipment manufacturing</td>
<td>CNC machine operational jobs</td>
<td>1</td>
</tr>
<tr>
<td>general logistics</td>
<td>order picking</td>
<td>3</td>
</tr>
<tr>
<td>general production</td>
<td>CNC machine operational jobs</td>
<td>1</td>
</tr>
<tr>
<td>general production</td>
<td>education and training</td>
<td>1</td>
</tr>
<tr>
<td>general production</td>
<td>general assembly</td>
<td>4</td>
</tr>
<tr>
<td>general production</td>
<td>general manufacturing</td>
<td>3</td>
</tr>
<tr>
<td>general production</td>
<td>general operations</td>
<td>1</td>
</tr>
<tr>
<td>general production</td>
<td>other, general</td>
<td>1</td>
</tr>
<tr>
<td>other, general</td>
<td>education and training</td>
<td>1</td>
</tr>
<tr>
<td>other, general</td>
<td>other, general</td>
<td>2</td>
</tr>
<tr>
<td>sheltered workplace</td>
<td>general assembly</td>
<td>5</td>
</tr>
<tr>
<td>warehouse</td>
<td>order picking</td>
<td>2</td>
</tr>
<tr>
<td>warehouse</td>
<td>other, general</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 108. Line graphs depicting each cycle time for each participant throughout the study.
List of My Publications


