Proceedings of the Human Factors and Ergonomics Society Annual Meeting
2023, Vol. 67(1) 104–110

DOI: 10.1177/21695067231196239
journals.sagepub.com/home/pro
Sage

Evaluating the Consistency of Cooperative Video Games in Inducing Teamwork Behaviors

Yvonne A. Farah[1], Antonio Banuelos-Moriel[1], and Michael C. Dorneich[1]

Abstract

Teamwork assessment requires research environments that replicate the task complexity and induce challenges and uncertainty. Gamified environments are gaining attention as suitable environments for teamwork assessment, where teaming behaviors can be quantified. However, further research is needed to establish the relationship between gaming mechanics and teaming competencies. This work aims to assess the consistency of video games in inducing teaming behaviors. Five video games were analyzed through coding behaviors and cooperative features from gameplay streams. Results show that the selected games induce similar behavioral profiles of percentages of teaming competencies. These games have similar design characteristics, including performance environments and game loops. Design suggestions reflect potential associations between the analyzed game mechanics and the resulting behavioral profiles. The resulting teamwork profiles can be targeted through the simulation cooperative genre, therefore providing a design tool for teamwork testbed designers to induce teamwork distributions of interest.

Keywords

Teamwork Assessment, Cooperative Video Games, Teamwork Testbeds, Teamwork Measurement, Cooperative Features, Behavioral Markers

Introduction

Teams operate in increasingly complex environments (Salas et al., 2005). Research environments are needed to assess the efficiency of teaming behaviors. Effective team assessment needs task environments, challenging events, measures of team performance, and competency metrics (Cooke et al., 2019). Gamified environments meet these requirements by providing dynamic gaming contexts where teaming metrics can be quantified and validated (Burgess et al., 2019).

Despite the expanding use of commercial and serious games, previous work suggests that using video games for teamwork assessment requires further empirical support for their validity (Mayer, 2018). Marlow et al. (2016) emphasized the current literature gap concerning a need for an understanding of how game mechanics induce teaming behaviors.

This work aims to assess the consistency of video games in inducing teaming competencies. Cooperative gameplay footage from publicly available streams (e.g., YouTube) was analyzed using a developed codebook to identify teamwork behaviors and the associated game features.

Background

Teamwork behaviors in gamified environments are informed by research in teamwork measurement systems and teamwork assessment in cooperative video games.

Teamwork Measurement Systems

Teams are groups of two or more individuals performing interdependent tasks toward a common goal (Salas et al., 1992). A team measurement system aims to capture team processes and collect team inputs and outcomes. The Input-Process-Output (IPO) model has been adapted to guide the development of measurement techniques (Wiese et al., 2015; Ilgen et al., 2005). The Process step assesses the behavioral, mental, and cognitive processes and emergent states. Team

[1]Iowa State University, Ames, IA, USA

Corresponding Author:
Yvonne A. Farah, Iowa State University, 2529 Union Dr., Ames, IA 50011-2140, USA.
Email: yfarah@iastate.edu

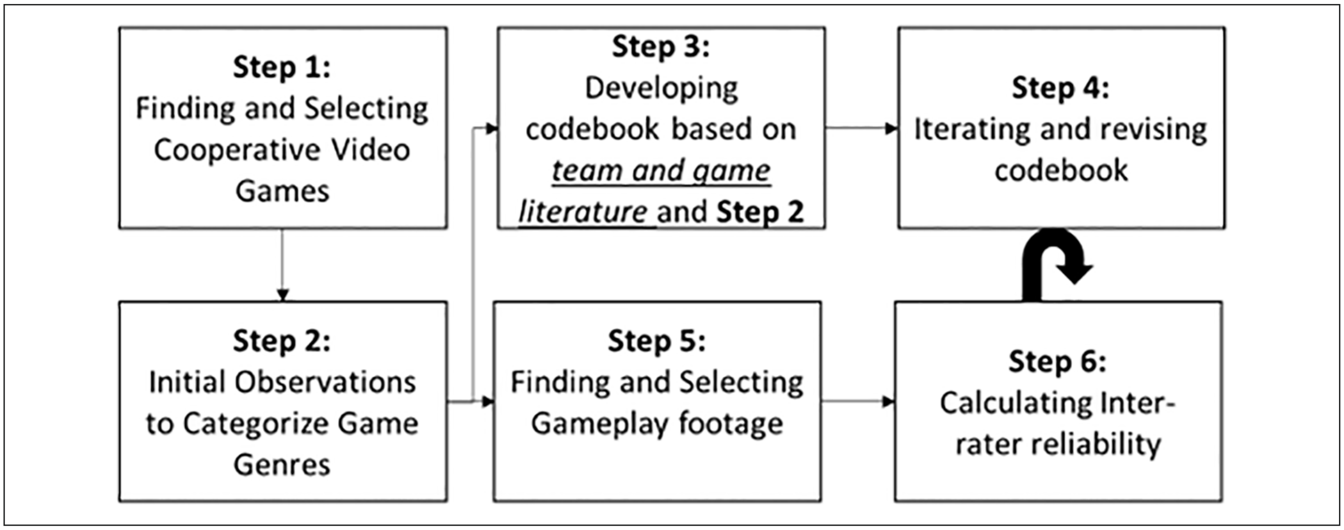

Figure 1. A diagram showing the phases of game selection and coding.

members can be asked to access their cognition, but self-assessment (e.g., surveys, interviews) can be subjective and biased (Wiese et al., 2015; Sottilare et al., 2018). Independent raters, trained to understand and track the assessed behaviors, can measure team skills. Examples of such systems include behaviorally anchored rating scales (BARS) and behavioral observational scales (BOS) (Kendall & Salas, 2004). Using these techniques involves developing behavioral markers, which are observable indicators of the existence or absence of a teaming competency (Rosen et al., 2011). Behavioral markers do not rely on self-report measures (Sottilare et al., 2018) but require researchers to develop concrete definitions of tracked and observable teaming behaviors.

Teamwork in Cooperative Video Games

In previous work, the authors started investigating the associations between cooperative game features and teaming behaviors (Farah et al., 2022), building on empirical and theoretical support from the literature. Depping & Mandryk (2017) used cooperation in design by implementing shared goals and interdependence through the degree to which team members rely on each other. Rocha et al. (2008) provided a list of cooperative game mechanics, and Depping & Mandryk (2017) associated some of them (e.g., complementary abilities, abilities that affect other players) as examples of mechanics that enhance interdependence. They reported that cooperation and interdependence had a significant main effect on experienced in-game relatedness and enjoyment. However, the work did not implement a team measurement system approach but focused on social interactions.

Different levels of interdependence and cooperative game mechanics were also investigated in asymmetric games, where two players collaborate but have different information, abilities, and interfaces (Harris, 2019). Interdependence had a significant main effect on connectedness and behavioral engagement (Harris, 2019).

To assess serious games as suitable environments for team research, Mayer (2018) developed TEAMUP, with various shared puzzles. They reported that gaming expertise, team efficacy, and cohesion were significantly associated with performance metrics such as scores and errors. These findings suggest the internal validity of using video games for teamwork assessment.

This work aims to build on previous work by investigating how consistently the mechanics in cooperative games can induce teaming behaviors. Suppose the nature of the relationship between teaming competencies and cooperative game mechanics can be established. In that case, the results of this work can be used to develop design suggestions to guide researchers and game designers.

Method

Figure 1 describes selecting and categorizing video games, selecting gameplay footage, developing, and iterating the codebook, and calculating inter-rater reliability.

Video Game Selection and Codebook development

The researchers used web search keywords: cooperative, teamwork, and multiplayer video games to select video games. Inclusion criteria included three characteristics (Morchheuser et al., 2017):

- Cooperative Video Game (support shared goals)
- Rules and mechanics that encourage interdependence (Seif El-Nasr et al., 2011; Rocha et al., 2008)
- Support verbal communication between players.

Table 1. Video Games' Descriptions.

Video Game	Description	Acronym
Overcooked 2 (Team 17, 2020)	Cooking simulation. Players navigate the kitchen, prepare and cook ingredients, and assemble and deliver orders.	**OC2**
Unrailed! (Indoor Astronaut, 2019)	Building train track. Players collect resources and mine obstacles to build tracks and avoid crashing.	**UR**
KeyWe (Stonewheat & Sons, 2020)	Assembling and sending letters. Players gather words, type letters, and deliver packages.	**KW**
Lovers in a Dangerous Spacetime (Asteroid Base, 2015)	Multiplayer space shooter. Players occupy a spaceship where they navigate the environment to defeat enemies, rescue friends and avoid crashing.	**LDST**
Catastronauts (Intertia Game Studios, 2018)	Multiplayer space-ship action. Players occupy a spaceship to work on different tasks (repair, shoot enemy, deliver tools) to defeat the enemy and avoid crashing.	**CN**

Researchers compiled a list of 53 video games as initial candidates found through web searches (Step 1). Researchers did an initial observation (Step 2) to assess the inclusion criteria through open coding (taking general notes of observations, mechanics, and observed behaviors). Some video games were dropped from the list due to the limited availability of videos on YouTube. The codebook was developed in parallel with this process (Step 3). Farah et al. (2022) describe a portion of the codebook. It went through several iterations where four people coded a video and discussed discrepancies. Two coders calculated inter-rater reliability (Step 6) and refined the codebook (Step 4) in the final iteration. One coder proceeded to code the final selection of gameplay footage using the final version of the codebook. The gameplay videos were selected (Step 5) by following three inclusion criteria:

- Clear commentary and distinguishable player voices
- Player characters are consistently observable on screen through a split or shared screen.
- Gameplay footage available for the targeted levels or amount of time (varies by genre)

Simulation Genre

This study reports the findings from analyzing five video games categorized under the simulation genre. The genre is characterized by the following:

- Players perform individual tasks to support shared goals.
- Interdependence is induced through a shared performance environment (e.g., assembling a dish, assembling a train track, navigating a ship, assembling a letter)
- Players engage in a game loop (levels have the same actions repeated over and over, e.g., assembling the same dish over several minutes and navigating a spaceship through the whole level).

Table 2. Number of Teams, Levels, and Average Time per Video Game.

Video Game	Teams	Levels	Average Time
OC2	10	Level 1-1 to 1-5	15
UR	10	4 to 8 levels	23
KW	10	Levels 1 to 4	13
LDST	8	Level 1	10
CN	9	Level 1-1 to 1-5	15

Table 1 describes the video games that were analyzed.

Table 2 summarizes the number of teams analyzed per video game, the levels, and the average time in minutes per video game.

Inter-rater Reliability

Inter-rater reliability (IRR) can present a way to ensure consistency across multiple coders and apply codebooks in qualitative research (McDonald et al., 2019). Percent or proportion agreement is a basic inter-rater agreement index that calculates a ratio between the number of agreements over the total number of codes (Gisev et al., 2013). While it does not account for the effect of chance in achieving agreement among raters, it is still used and reported (Gisev et al., 2013). Values between 41-60% are considered moderate, and 61-80% are considered substantial.

During the codebook refinement phase, two coders would code 10 minutes of the same video. If the IRR were less than 75%, the two researchers would meet to discuss the coding, refine their mutual understanding, and update the codebook. This process was completed for two more iterations to refine the final codebook definitions. The final IRR values are 73% for Overcooked2, 73% for Lovers in a Dangerous Spacetime, and 72% for Key We. After refining the codebook, the final version was used by one coder to code the selected gameplay footage.

Table 3. Descriptions of Tracked Teaming Competencies.

Competency	Description
Situation Assessment	Developing a shared understanding of the environment (identifying relevant cues and communicating meaning) (Rosen et al., 2011).
Analysis and Planning	- Interpreting and evaluating teams' tasks and available resources. - Developing a course of action based on the team's discussion (Rosen et al., 2011; Marks et al., 2001)
Explicit Coordination	- Explicitly sequencing or synchronizing the team's actions by verbally assigning sequential tasks (Rosen et al., 2011) - Reporting status, location, and needs to the team. (Rosen et al., 2011)
Implicit Coordination	- Sequencing or synchronizing the team's actions without verbal communication. (Espinosa et al., 2004) - Doing actions that contribute to the team's progress without verbal communication.
Monitoring and Backup	- Monitoring team members' progress (Salas et al., 2005). - Providing feedback, resources, or behavioral help when an imbalanced workload is noticed (Salas et al., 2005).
Systems Monitoring	- Internal tracking of team resources (Marks et al., 2001). - External tracking of environmental conditions (Marks et al., 2001).
Team Cohesion and Social	- Verbally expressing encouragement or complementing team skills or abilities (Sottilare et al., 2018). - Engaging in social conversations inspired by games' themes.
Adaptive Behaviors	- Diagnosing and evaluating performance through reflection and learning (Rosen et al., 2011). - Developing a new strategy resulting from failure in performance (Rosen et al., 2011).

Results

Team Codes

Table 3 summarizes the teaming competency categories. Every competency compiles several relevant behavioral markers. For example, for situation assessment, two behaviors were coded: 1) recognizing cues in the environment (e.g., noticing a button, noticing instructions, recognizing a puzzle component), and 2) communicating the meaning of the cue (e.g., I think the rice (cue) needs to be cooked (meaning). If I move the joystick (cue), the crane moves (meaning).

Consistency

Table 4 presents the average percentages of every teaming competency and the standard error. For every competency, the percentage represents the total frequencies of behavioral markers associated with the competency, divided by the total number of codes. Percentages were derived for every team, and the table presents the teams' average and standard error per video game.

Figure 2 shows the pattern of each of the five video games across the teaming competencies.

Pattern Similarity. To assess the similarity between the five video games in inducing similar behavioral teaming profiles, each video game was considered a vector of values representing percentages of every competency. The cosine similarity calculates the cosine of the angle between two vectors as a measure of the similarity between them (Cha, 2007; Han et al., 2012). It is a value ranging between 0 and 1,

Table 4. Average Percentages and Standard Errors per Teaming Competency per Video Game.

Competency	KW	OC2	UR	CN	LDST
Situation Assessment	20 (1)	18 (1)	6 (1)	9 (1)	11 (3)
Analysis and Planning	4 (1)	5 (1)	11 (1)	11 (1)	5 (2)
Explicit Coordination	24 (2)	24 (1)	23 (2)	24 (1)	18 (1)
Implicit Coordination	21 (3)	27 (2)	31 (3)	21 (1)	31 (2)
Monitoring and Backup	6 (1)	8 (1)	6 (1)	9 (1)	5 (1)
Systems Monitoring	22 (1)	13 (1)	20 (1)	23 (1)	29 (3)
Cohesion and Social	1 (0)	3 (1)	2 (0)	2 (0)	0 (0)
Adaptive Behaviors	2 (0)	1 (0)	1 (0)	1 (0)	1 (0)

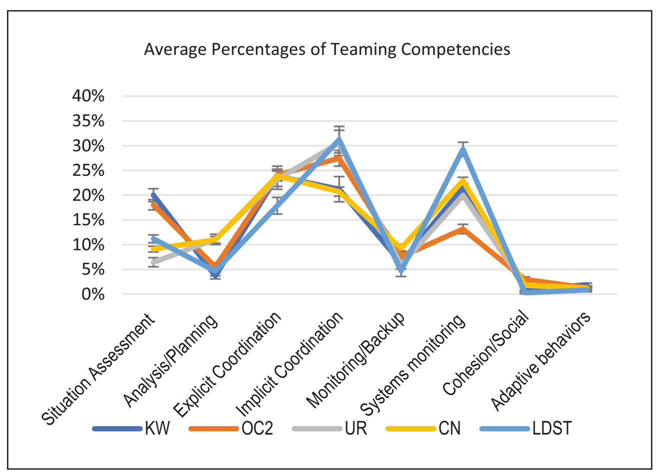

Figure 2. Line charts representing percentages of teaming competencies for every game.

Table 5. Cosine Similarity Values.

	OC2	UR	CN	LDST
KW	0.97	0.92	0.95	0.94
OC2		0.94	0.93	0.92
UR			0.97	0.96
CN				0.95

determining whether two vectors follow the same direction. Table 5 shows the cosine similarity values between the five-game vectors. Overall, the average Cosine similarity was 0.94.

Discussion

Similarity and Consistency

Some competencies were consistently observed with higher frequencies in the line chart: explicit and implicit coordination were consistently ranked in the top three. Systems monitoring was also observed among the top competencies. Other teaming processes such as situation assessment, analysis, planning, monitoring, and backup were observed with low to intermediate percentages, while cohesion, social and adaptive behaviors were observed with low percentages. The similar percentages of these competencies induced by the five video games suggest consistency in the behavioral profiles.

The cosine similarity measure assessed the angle between the five-game vectors with an average of 0.94. The closer the cosine value is to 1, the lower the angle between the vectors; therefore, they are more similar in direction. Across all game pairs, the cosine was higher than 0.9, showing a high similarity between games. The following sections discuss the implications of these findings.

Competencies with top percentages: Coordination and Systems Monitoring

Explicit Coordination, Implicit Coordination, and Systems Monitoring peak frequently across the analyzed games appearing in the top three percentages for KeyWe, Unrailed, Catastronauts, and Lovers in a Dangerous Spacetime. Coordination has been defined as managing dependencies across tasks, resources, and people (Malone & Corwston, 1990). In gamified environments, dependencies can be optional or mandatory, and they arise from cooperative mechanics that encourage players to rely on each other (Harris, 2019). The analyzed games had a variety of interdependencies. Players engaged in multi-directional sequential tasks (Saavedra et al., 1993) (e.g., following specific sequential orders in Overcooked 2 and KeyWe). A simultaneous, multi-directional workflow was also present (Saavedra et al., 1993) (e.g., collecting wood for the train track while mining obstacles and assembling tracks in Unrailed or shooting enemies while repairing the ship and sharing tools in Catastronauts). Players were required to manage these dependencies spreading across their tasks, resources, and team players through sequencing and synchronizing their actions and constantly reporting their progress, needs, and status.

Previous work suggests that team members must form and maintain the team's awareness of multiplayer video games by gathering information from the environment (Wuertz et al., 2018). The games provided players with rich and diverse external environments (enemies, resources, utilities) and internal environments (ship damage, food locations, task progress). Therefore, the coder could detect observable systems monitoring behaviors through the players' avatars (e.g., avatar looking around the environment), verbal communication (e.g., players reporting on incoming enemies, players reporting on train progress), or observable actions (e.g., shooting at incoming enemies, repairing ship when a crack appears, delivering food when a dish is fully assembled). Frequent verbal coordination resulted in a high percentage of systems monitoring since players support their shared environment awareness by communicating these observations to team members. Through observations, the coder did notice frequent associations between systems monitoring and verbal reporting (e.g., players noticing a fire spreading in the spaceship and reporting it).

Simulation games are characterized by each level using the same gameplay loop, a cycle of repeated actions in a game. Teams could execute interdependent tasks through implicit coordination (e.g., assembling a dish in a specific order, shooting at enemies simultaneously when detected, getting, assembling tracks, and typing letters in a word). Previous work suggests that expert teams engage in implicit coordination and depend less on verbal communication (Cannon-Bowers et al., 1993; Wuertz et al., 2018). While this paper does not compare experts and novices, all teams (regardless of their expertise level) eventually became familiar with the game mechanic through the game loop and therefore became more capable of performing implicitly.

Situation Assessment and Analysis

Teamwork behaviors go through time-based cycles. Marks et al. (2001) suggested a time-based teaming framework where teams move across transition behaviors, such as situation assessment and analysis, and execution behaviors, such as coordination and planning. They proposed that depending on the task demands, teams can spend more time in certain stages (transition or execution) and alternate between them more or less frequently. The coded behavior markers related to assessment and analysis focused on the teams' behaviors of noticing mission-relevant cues, communicating their meaning, and analyzing or discussing a way to solve the mission and develop a course of action. In KeyWe and Overcooked, situation assessment ranked higher compared to other simulation games. In both games, players spend several minutes in one game level (e.g., assembling sushi for three minutes or typing and delivering letters) and then move to a new level. With every new level, players must identify new cues and figure out how to solve the puzzle before engaging in execution. Therefore, due to exposure to new levels, situation assessment behaviors were observed with higher percentages in these two games.

Conclusion

This paper analyzes five cooperative video games from the simulation genre. The results show a highly consistent pattern of behavioral profiles across the five games, demonstrating that video games with similar cooperative mechanics and performance environments induce consistent behavioral profiles. The cosine similarity measure was 0.94, suggesting a high similarity in the pattern. Specifically, the analyzed genre was associated with explicit coordination, implicit coordination, and systems monitoring.

The work has several limitations. While the codebook was developed from theoretical teaming research and iterated through coding and discussions, there remains a level of subjectivity or potential inconsistencies associated with human coding. Future work will look to develop automated ways of coding this type of data.

Secondly, the data was collected from publicly available footage, and therefore the researchers need direct access to the teams' input levels, such as expertise and familiarity. This can create limitations when it comes to analyzing teams with different competencies. Yet, the analyzed footage and games reflect quasi-naturalistic settings where gamers perform outside the lab environments. This is an encouraged method in games' user research since it allows for natural emergent states (Harris, 2019). However, developing a gamified testbed where researchers control the simulation parameters can provide more opportunities to test different mechanics and induce targeted teaming behaviors.

In conclusion, through this analysis, the researchers aim to establish cooperative games as environments that can systematically induce behavioral profiles relevant to the implemented cooperative mechanics.

Acknowledgments

This project was funded in part by the Game2Work research program as part of the Iowa State University Presidential Interdisciplinary Research Initiative.

This project is based upon work supported in part by the National Science Foundation under Award No. 1757900. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Authors would like to thank: Eli Whitney, Emma Edgar, and Thomas Walker for contributing to the codebook iteration process.

ORCID iD

Michael C. Dorneich https://orcid.org/0000-0001-6386-4787

References

Asteroid Base. (2015). Lovers in a Dangerous Spacetime [Video Game].

Bishop, J., Burgess, J., Ramos, C., Driggs, J. B., Williams, T., Tossell, C. C., . . . de Visser, E. J. (2020, April). Chaopt: a testbed for evaluating human-autonomy team collaboration using the video game overcooked! 2. In *2020 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 1-6). IEEE.

Burke, C. S., Stagl, K. C., Salas, E., Pierce, L., & Kendall, D. (2006). Understanding team adaptation: a conceptual analysis and model. *Journal of Applied Psychology*, *91*(6), 1189.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision-making.

Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, *1*(2), 1.

Cooke, N., Demir, M., & Huang, L. (2020). A framework for human-autonomy team research. In *Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22* (pp. 134-146). Springer International Publishing.

Depping, A. E., & Mandryk, R. L. (2017, October). Cooperation and interdependence: How multiplayer games increase social closeness. In *Proceedings of the Annual Symposium on computer-human interaction in play* (pp. 449-461).

Espinosa, J. A., Lerch, F. J., & Kraut, R. E. (2004). Explicit versus implicit coordination mechanisms and task dependencies: One size does not fit all.

Farah, Y. A., Dorneich, M. C., & Gilbert, S. B. (2022, September). Evaluating Team Metrics in Cooperative Video Games. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 66, No. 1, pp. 70-74). Sage CA: Los Angeles, CA: SAGE Publications.

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338.

Harris, J. J. (2019). Leveraging asymmetry and interdependence to enhance social connectedness in cooperative digital games.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annu. Rev. Psychol.*, *56*, 517-543.

Indoor Astronaut. 2019. Unrailed! [Video Game]. Daedalic Entertainment GmbH, Bilibili Co., Ltd.

Interia Game Studios. 2018. Catastronauts [Video Game].

Kendall, D. L., & Salas, E. (2004). Measuring team performance: Review of current methods and consideration of future needs. *The science and simulation of human performance*, *5*, 307-326.

Malone, T. W., & Crowston, K. (1990, September). What is coordination theory, and how can it help design cooperative work systems? In *Proceedings of the 1990 ACM conference on Computer-supported cooperative work* (pp. 357-370).

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of management review*, *26*(3), 356-376.

Marlow, S. L., Salas, E., Landon, L. B., & Presnell, B. (2016). Eliciting teamwork with game attributes: A systematic review and research agenda. *Computers in Human Behavior*, *55*, 413-423.

Mayer, I., (2018). Assessment of teams in a digital game environment. *Simulation & Gaming*, *49*(6), 602-619.

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction*, *3*(CSCW), 1-23.

Morschheuser, B., Riar, M., Hamari, J., & Maedche, A. (2017). How do games induce cooperation? A study on the relationship between game features and we-intentions in an augmented reality game. *Computers in human behavior*, *77*, 169-183.

Rocha, J. B., Mascarenhas, S., & Prada, R. (2008). Game mechanics for cooperative games. *ZON Digital Games 2008*, 72-80.

Rosen, M. A., Bedwell, W. L., Wildman, J. L., Fritzsche, B. A., Salas, E., & Burke, C. S. (2011). Managing adaptive performance in teams: Guiding principles and behavioral markers for measurement. *Human resource management review*, *21*(2), 107-122.

Saavedra, R., Earley, P. C., & Van Dyne, L. (1993). Complex interdependence in task-performing groups. *Journal of applied psychology*, *78*(1), 61.

Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training.

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a “big five” in teamwork? *Small group research*, *36*(5), 555-599.

Seif El-Nasr, M., Aghabeigi, B., Milam, D., Erfani, M., Lameman, B., Maygoli, H., & Mah, S. (2010, April). Understanding and evaluating cooperative games. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 253-262).

Sottilare, R. A., Shawn Burke, C., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2018). Designing adaptive instruction for teams: A meta-analysis. *International Journal of Artificial Intelligence in Education*, *28*, 225-264.

Stonewheat & Sons. 2021. KeyWe [Video Game]. Fireshine Games.

Team 17, Ghost Town Games. 2018. Overcooked2! [Video Game].

Wiese, C. W., Shuffler, M. L., & Salas, E. (2015). Teamwork and team performance measurement.

Wuertz, J., Alharthi, S. A., Hamilton, W. A., Bateman, S., Gutwin, C., Tang, A., . . . Hammer, J. (2018, April). A design framework for awareness cues in distributed multiplayer games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-14)