Teamwork and Adaptation in Games (TAG): A Survey to Gauge Teamwork

Casper Hartevelde
Northeastern University
c.hartevelde@northeastern.edu

Dylan Schouten
Northeastern University
dylan.schouten@gmail.com

Wade Kimbrough
Extra Ludic Inc.
wade@extraludic.com

Erica Kleinman
Northeastern University
erikleinman@gmail.com

Truong Huy Nguyen
Fordham University
truonghuy@gmail.com

Paul Fombelle
Northeastern University
p.fombelle@northeastern.edu

Paola Rizzo
Northeastern University
p.rizzo@northeastern.edu

Samuel Liberty
Extra Ludic Inc.
samuel.a.liberty@gmail.com

Magy Seif El-Nasr
Northeastern University
magy@neu.edu

ABSTRACT

Teamwork is becoming an important element of gameplay. Many games nowadays incorporate teamwork in one form or another. As a result, many game researchers have investigated aspects of teamwork. However, we have yet to develop effective surveys to gauge how people strategize or problem solve together within different games, and how such games enable these activities. In this paper, we propose the *Teamwork and Adaptation in Games* (TAG) survey. We developed this survey using theories of team adaptation, a process of how teams work together and adapt their strategies to changes in environments. Through three separate studies, the TAG survey was used to gauge team adaptation. Our results show that the TAG survey is an effective instrument allowing us to gauge team adaptation in different game environments. We further found that the instrument can be effectively used to compare games.

CCS CONCEPTS

- **General and reference** → **Measurement**; *Surveys and overviews*;
- **Human-centered computing** → *User studies*; *Field studies*; Computer supported cooperative work.

KEYWORDS

survey design, teamwork, adaptation, digital games

ACM Reference Format:

Casper Hartevelde, Erica Kleinman, Paola Rizzo, Dylan Schouten, Truong Huy Nguyen, Samuel Liberty, Wade Kimbrough, Paul Fombelle, and Magy Seif El-Nasr. 2019. Teamwork and Adaptation in Games (TAG): A Survey to Gauge Teamwork. In *The Fourteenth International Conference on the Foundations of Digital Games (FDG '19)*, August 26–30, 2019, San Luis Obispo, CA, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3337722.3337731>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FDG '19, August 26–30, 2019, San Luis Obispo, CA, USA

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-7217-6/19/08...\$15.00

<https://doi.org/10.1145/3337722.3337731>

1 INTRODUCTION

Team-based gaming has become an important avenue of game design. Teamwork manifests differently across games: Massively Multiplayer Online (MMO) games, such as *World of Warcraft* and *EVE Online*, have players self-organize into long-term power structures in order to gain influence and tackle the game's largest challenges [27, 34], while Multiplayer Online Battle Arena (MOBA) games, such as *League of Legends* and *Dota 2*, put players in game-composed short-term teams that square off in formalized matches [11, 23, 53]. With the rise of E-Sports, an industry with revenues reaching a billion USD in 2018 [42], teamwork will become a more prevalent aspect of play. However, research into how teams strategize, problem-solve, and adapt to the dynamic game environments is still in its infancy: Most of the existing works focus on opportunities for training or identifying factors of success in gameplay [21, 27, 39, 40] or understanding mechanics for co-operative play [47, 50] rather than focusing on team dynamics and team processes, including teamwork and adaptability. Given the importance and prevalence of teamwork in games and the current gap in understanding this phenomenon, further research is needed. Such research will have multiple utilities, including understanding cooperative work, highlighting the differences between mechanics used to support teamwork, which would provide valuable insight for game designers seeking to enhance teamwork in their games [e.g., 1, 2].

To address this gap, we propose a new survey instrument to capture teamwork and adaptation within games. While there has not been work looking into team strategy and problem solving processes, there has been some work studying teams in games by applying data driven analysis and modeling [21, 36, 38, 39, 46, 54] or qualitative methods [3, 20, 23, 35, 40, 52]. We propose a self-report approach. Surveys are easily scalable, particularly online surveys (cf., [14]), allowing us to gauge teamwork at an operational scale and capturing the variance expected due to the contextual and individual differences. Additionally, self-report measures can combine open and closed questions to measure a wide range of cognitive variables and processes related to teamwork and adaptation, such as situational assessment, strategy formulation, and team learning. Closed questions provide immediately actionable data for follow-up analysis, while open questions can be used to characterize team processes or enable in-depth qualitative analyses.

The purpose of our work is two-fold: (1) develop a novel survey instrument for measuring teamwork and adaptation processes within games, and (2) demonstrate the utility of this instrument. For the first goal we developed the “Teamwork and Adaptation in Games” (TAG) survey composed of closed questions intended to quantify adaptation processes and open questions intended to understand these processes in specific gaming contexts. We based our survey on Rosen et al.’s [43] four-factor model (described in more detail in Section 2). In this paper, we focus on the survey’s closed questions, keeping the analysis of the open questions for future work. We conducted exploratory and confirmatory factor analyses using data from three different studies comprising 415 players to refine the survey and determine its underlying theoretical model. For the second goal we investigated the utility of TAG as a method to understand teamwork in two gaming environments: *MarketPlace Live* (MPL), a business simulation game used in university economics and marketing classes [6], and *Daedalus*, an escape room-style Alternate Reality Game (ARG) wherein teams of players compete for prize money. Through these two goals we present the TAG survey as a contribution and demonstrate its utility in quantifying and understanding the team adaptation process. As a byproduct, we show how the TAG is useful in comparing game environments in terms of their affordances to foster teamwork.

2 BACKGROUND

To accomplish our aims, we delved into the general theory of teamwork to see what we could learn from this domain and apply to games. More specifically, we considered the existing methodologies for studying teamwork. This investigation led us to conclude that we had to develop our own instrument. First we will detail what is known about teamwork in games and why we pursued this effort.

2.1 Teamwork in Games

Since teamwork occurs in a wide variety of games, from board games to local party games to multiplayer online games, such as (E-)sports, much research has been devoted to understanding how teams form and function in these types of games as they present unique opportunities for collaboration and social interaction among players [27, 35]. Aside from studying them to better understand teamwork *in* games, some research uses these environments to gain insight into how people organize themselves [40, 51], and how to improve teamwork environments in the real world [27, 40].

Research efforts can roughly be divided into two methodological camps. The first camp is quantitative and focuses primarily on data-driven analysis and modeling [21]. For example, Pobiedina et al. [38, 39] used a data-driven analysis of team gameplay data to identify success prediction facets in the MOBA game *Dota 2*. This type of work requires access to large data sets and careful understanding of game data structures. Additionally, as it is restricted to behavioral data, researchers are limited in their ability to infer how teams plan, strategize, and organize themselves.

The second camp is qualitative. Some researchers make use of game chat transcripts, open-ended interviews, and audiovisual communication data to make sense of teamwork [20, 23, 52], while others use *in-situ* participant observation of co-located or distributed team members [3, 35]. For example, Prax [40] shows that leaders of

raiding guilds in *World of Warcraft* regulate their leadership styles between a results-focused approach (during raids) and a people-focused approach (during regular gameplay). While this approach can provide deeper insight into understanding teamwork processes, in all circumstances it requires dozens or hundreds of work hours to transcribe, code, and interpret for meaning. Therefore, it typically focuses on small numbers, thereby limiting any theoretical claims.

In our work we seek to complement both camps through developing a survey instrument. To our knowledge, there are no surveys on teamwork and adaptation in the context of a game environment. We aimed to ground our development of the instrument using existing teamwork literature, which we discuss next.

2.2 Theory of Teamwork

Salas et al. define a team as “*two or more people who interact dynamically, interdependently, and adaptively towards a common goal / object / mission, who have been assigned specific roles or functions to perform and who have a limited life span of membership*” [44, p. 4]. Teamwork can then be seen as the set of interactions and activities team members engage in to achieve a common goal. The literature discusses many important constructs that determine how teams perform, such as situation assessment [13, 45], (shared) mental models [7, 30], trust [9, 33], communication [49], and team composition [31], and important processes, such as briefing/debriefing [48], planning [49], and training [18].

For our work, we focused on the concept of *adaptation*, which Burke et al. [5] describe as those processes that teams use to adjust their performance for better outcomes. We opted for this concept, because games are dynamic environments, and thus adaptation is an important element for all team based games, and there is little existing work on teamwork in games. To measure adaptation, we primarily adhere to Rosen et al.’s [5, 43] process model of adaptation, which consists of the following four concurrent phases:

- (1) *Situation assessment (SA)*: the process of information gathering where the team searches for cues relevant to the mission. It encompasses key sub-processes of: cue recognition, meaning ascription, and team communication.
- (2) *Plan formulation (PF)*: involves designing a sequence of actions to transform the current situation into a desired one, and consists of the sub processes of: mission analysis, goal specification, deliberate and contingency planning, role differentiation, and preemptive conflict management.
- (3) *Plan execution (PE)*: team members engage in behaviors to execute the plan designed during the plan formulation phase. It encompasses: coordination, mutual monitoring, back-up behavior, systems monitoring, reactive strategy planning, reactive conflict management, and affect management.
- (4) *Team learning/reflection (TL)*: the process by which changes emerge in the team’s behavior as a result of the continuous acquisition of knowledge and performance feedback, and encompasses reflection and critique processes in a number of contexts including the recapping of events and the assessment of strengths and weaknesses.

The Rosen et al. [43] model combines a systematic view of teamwork processes with a broad overview of relevant behavior markers. This made it a good candidate for developing our survey.

2.3 Methodologies for Studying Teamwork

As per the previous section, we focused on measuring adaptation and followed Rosen et al.'s [43] work to define a process by which we can measure SA, PF, PE, and TL. However, the behavior markers from this model are typically observed by either videotaping the interaction and having experts code the interaction for these measures, or by managers observing their employees and noting these behaviors [12]. As we intended to develop a survey, we sought existing literature that could provide guidance on measuring these concepts. Additionally, because we approach team process as a dynamic process [24], we were ideally looking for a survey instrument that would allow for multiple measurements at various points of team interaction, and thus we needed a survey that is constrained in length, minimally obtrusive, and requires minimal effort to fill.

Research on teamwork and team processes has yielded several novel methodologies. For example, for measuring situation awareness, proposed instruments include ones that: (1) ask participants for a judgment on the number of dimensions related to Situation Awareness (SART); (2) show participants all information and measuring latency instead of correct answers (SPAM); and (3) measure the speed and accuracy by which an error is detected and corrected [8, 12]. However, aside that these instruments do not always fit in the context of a survey, they do not satisfy our constraints. The only survey we found within this search was one on adaptability which specifically aims to understand adaptation as a trait of an individual rather than a team process [37, 55]. This survey was also not validated or widely applied. For this reason, we decided to develop our own instrument.

3 TAG SURVEY

We developed the TAG survey in response to the critical need for an instrument to measure teamwork and adaptation in games, as discussed in the previous section. In designing the TAG, we aimed for a relatively short survey that can be deployed multiple times and measure changes in teamwork across an experience while not disrupting the game flow. Although the survey can be used on its own, such a survey would also be very suitable in broader approaches, in particular those that include data-driven techniques to measure behavior. The TAG survey uses a mix of open and closed questions, both based on Rosen et al.'s [43] model of adaptation. Table 1 shows all questions of the most current TAG version and the excluded questions from the first version.

To design the closed questions, we converted the “behavioral markers” for each phase of Rosen et al., [43, p. 115] into questions targeting that phase in the context of playing a game; the questions are modified per game, where <game name> is replaced by the actual game name used for the study. To illustrate how such markers were converted to questions, we take the Plan Formulation phase as an example. This phase includes “*Explicitly articulate the team’s objectives*” as a behavioral marker. We transformed this into the question “*My team explicitly articulated its objectives.*” Between 6 and 10 questions were created for each of the four phases in this manner, resulting in version 1 composed of 28 closed questions in total. In this process, behavioral markers that were difficult to self-report on and realistically only measurable through direct observation were ignored. The questions were written on a 5-point

bipolar Likert scale ranging from Strongly Disagree to Strongly Agree. The choice for this scale is rooted in the familiarity and ease that participants would have in filling out the survey, all which would help facilitate creating an instrument that is minimally obtrusive and easily deployed. All closed questions are positively phrased, except for TAGC.17, which we felt was best written in this manner to avoid a positive response bias on this sensitive topic. Although there is some disagreement about the need to include reverse-coded questions in survey design [29], we chose to use such design, because reverse-coded questions are generally more difficult to answer and can equally lead to biased outcomes. After the first study (see Methods and Results section), based on factor analysis results and based on an expert review from Gallup, a company known for its opinion polling and survey design, we excluded three questions (TAGC.7, TAGC.20, and TAGC.27) and slightly reworded several questions (TAGC.3-5, 11-13, 15, 19, 21, and 22, and TAGO.1-5), to improve question clarity and reduce ambiguity.

To design the open questions, we used the Rosen et al. [43] behavioral markers that are more challenging to study on a Likert scale or through data because of the lack of game-based outcomes. For example, most teams engage in “*strategy formulation*” during face-to-face meetings, which does not result in observable game data. The team reflection process phase is also cognitive, making it impossible to measure through (behavioral) game data. The open questions were written to prompt respondents to elaborate on these processes in more detail, primarily focusing on the strategies and interaction methods used by their teams. Five open questions were created originally; a sixth question (TAGO.3) was added after analyzing the open responses, which did not give enough insight into the individual’s role, tasks, and responsibilities in a team. Note that the results from the open questions are not presented here.

The TAG is meant to be taken independently by each team member. Subscale scores can then be summarized by averaging the scores on each item per subscale, and per individual or group. Per group, scoring can be standardized by dividing the summarized scores by the number of team members. Table 1 shows the original subscales designed after Rosen et al.’s [43] model as well as new subscales defined after our factor analysis (see Results section).

4 METHODS

4.1 Participants

We conducted three studies with the TAG. We first administered the TAG with *MPL* in Fall 2017, targeting 96 teams (4–5 players each) and receiving 85 responses out of 391 players with an average of 3.29 people per team answering the survey. We then revised the survey based on the results and expert reviews, and administered it to a second set of *MPL* players in Spring 2018: 110 teams yielded 206 responses out of 462 players with an average of 5.80 people per team answering the survey. Finally, we applied the same revised TAG to *Daedalus* players in Fall 2018. We recruited 38 teams to play, 26 of which (for a total of 124 participants) filled out the TAG survey. These studies are from here on referred to as *MPL-1*, *MPL-2*, and *Daedalus*, respectively.

For the *MPL-1* and *MPL-2* studies, recruitment took place as part of an undergraduate marketing course, where playing *MPL* is

Label	Item	Rosen	TAG	Loading
TAGC.1	My team explicitly articulated its objectives	PF	GS	.59-.71
TAGC.2	My team brainstormed strategies on how to play <game name>	PF	GS	.67-.72
TAGC.3	My team members passed information to one another in an efficient manner	PE	GS	.59-.38
TAGC.4	My team identified potential problems it could encounter while playing <game name>	SA	GS	.60-.63
TAGC.5	My team created strategies for improving our performance in future <game stages>	TL	GS	.67-.76
TAGC.6	My team came up with a list of tasks to accomplish its objectives	PF	TP	.71-.76
TAGC.7	My team identified the root causes of conflicts between its members <i><deleted></i>	PE	<i>n/a</i>	<i>n/a</i>
TAGC.8	I asked my teammates for feedback regarding my performance	TL	FB	.82-.56
TAGC.9	My team referred to past strategies when making decisions	TL	GS	.69-.52
TAGC.10	My team brainstormed new plans and strategies when it encountered problems	PF	GS	.74-.58
TAGC.11	My team periodically updated a list of to-dos, needs, and objectives	PE	TP	.72-.67
TAGC.12	My team anticipated problems that could occur in the market	SA	GS	.63-.61
TAGC.13	I received feedback from my teammates about my performance	PE	FB	.92-.64
TAGC.14	I was assigned specific tasks according to my abilities	PF	TP	.57-.67
TAGC.15	My team tried to resolve team conflicts by negotiating	PE	GS	.57-.54
TAGC.16	My team articulated the best and worst strategies	PF	GS	.58-.72
TAGC.17	One or more of my teammates showed negative feelings such as stress, anger, depression, frustration or similar emotions that negatively impacted the team <i><reverse/deleted></i>	PE	<i>n/a</i>	<i>n/a</i>
TAGC.18	I analyzed team member contributions to identify their errors	TL	FB	.48-.63
TAGC.19	My team came up with a list of additional ideas in case a strategy did not work	PF	TP	.59-.64
TAGC.20	My team conversed without using sarcastic or non-constructive comments <i><deleted></i>	TL	<i>n/a</i>	<i>n/a</i>
TAGC.21	I provided feedback to teammates about their performance	TL	FB	.78-.67
TAGC.22	My team was able to see why our strategies were or were not successful thus far	SA	GS	.58-.52
TAGC.23	My team explicitly outlined everybody's responsibilities	PF	TP	.61-.69
TAGC.24	I accepted suggestions from teammates about how I could improve my performance	TL	FB	.67-.69
TAGC.25	My team identified the key issues for improving its performance in <game name>	TL	GS	.77-.79
TAGC.26	My team discussed errors and their causes	TL	GS	.72-.66
TAGC.27	My team attempted to provide opportunities for socializing <i><deleted></i>	PE	<i>n/a</i>	<i>n/a</i>
TAGC.28	When offered feedback, I was able to identify what would help me improve my performance	TL	FB	.67-.88
TAGO.1	What was your team's strategy for playing <game name>?			
TAGO.2	How did your team's strategies change during play? And why?			
TAGO.3	How would you describe your role in the team, and what were your tasks and responsibilities?			
TAGO.4	How did your team decide which actions to take?			
TAGO.5	What roles did your team assign to each of its members? Why? Did your team members fulfill this role and why yes or why not?			
TAGO.6	How did your team communicate?			

Table 1: The “Teamwork and Adaptation in Games” closed questions (TAGC) and open questions (TAGO). Original Rosen scales for closed questions: Situation Assessment (SA), Plan Formulation (PF), Plan Execution (PE), or Team Learning (TL). New TAG scales after factor analysis: Game Strategy (GS), Team Planning (TP), and Feedback (FB). The last column shows the standardized loadings resulting from the CFA with the final model for the MPL-2 and Daedalus study, respectively.

a required activity. The larger yield for MPL-2 is a result of our improved outreach efforts, and our direct access to students through direct contact with an instructor who teaches several sections. Student participants received remuneration for completing the TAG. For the Daedalus study, we engaged in a public recruitment effort through social media and word-of-mouth. Participants signed up as an entire team (consisting of 5–7 team members), a partial team (2–4 players), or as individuals. Importantly, every team member had to be: (1) 18+ years old, (2) a resident of the city the study was conducted in, and (3) a U.S. resident. The second requirement is needed as the game requires participants to go to physical locations, and the third is due to us offering cash prizes. Teams who finished

in the top five received \$300. Nevertheless, every team that finished received \$100, contingent on completing the TAG.

Note that teams composed of friends versus teams composed of strangers likely differ in the division of roles and teamwork as well as adaptability. However, studying this is beyond the scope of this study. In *MPL*, students were mostly asked to organize their own teams, and as such no composition data is available. Teams in *Daedalus* were mostly self-formed by groups of friends and/or acquaintances, sometimes with 1-3 strangers added to fill out their numbers; a few teams were made up of two separate, small groups of acquaintances, while a few others were made up completely of strangers.

4.2 Materials

We deployed the TAG in two different game environments, which we describe below. For MPL-1 we used the original TAG version with 28 questions and 5 open questions; for MPL-2 and Daedalus we used the revised and most current version with 25 questions and 6 open questions. For the analysis *R* version 3.4.2 was used with the main packages *psych* for traditional statistical analyses and *lavaan* and *semTools* for the confirmatory factor analyses.

MarketPlace Live (MPL) is an existing business/marketing simulation game designed for higher education, which has been in development and use for the past 25 years [6]. In this game, players work as a team and strategize to become influential first movers in a chosen emerging market (the version used here deals with 1990's-era personal computers). Players play the game in self-organized "company" teams. Over six "business quarters" (Q1–Q6), which last a length of time determined by the game administrator, each team attempts to increase their company's net profit and market share.

We partnered with independent game developer Extra Ludic Inc. to develop *Daedalus*, designed as an Alternate Reality Game (ARG) because of the potential for ARGs as research environment for social and behavioral science [17]. *Daedalus* revolves around a central narrative of a sentient AI testing players through a series of puzzles, in order to learn human problem-solving strategies so it can solve humanity's most pressing problems. Players enter a text-based, five-stage labyrinth of puzzles, which they need to solve to progress. The puzzles have been designed with the specific purpose of assessing aspects of team adaptability. Pulakos et al.'s [41] dimensions of adaptability were used as a guide in designing these puzzles. The game is delivered through the popular *Slack* online chat environment, and thus resembles early text-based games from the 1980s and examples of Interactive Fiction, while also providing a familiar "work" environment. A combination of Slack agents ("bots") and administrators directly chatting with players was used to organize the game, with a focus on the former: Administrators only directly talked to players at the start and end of the game, and in case unexpected errors occurred. Additionally, we used *ARIS*, a freely-available app designed to facilitate the creation of location-based games [25, 26], for integrating a physical component, as some puzzles require players to go into the real world for clues.

4.3 Procedure

MPL is deeply integrated into the semester-long marketing course. Students engage with the game through introductory course material, and then play on their own over several weeks. The only deadlines they are given is that each quarter of play must be completed at a set time. The instructors of each class are free to set their own deadlines and implement their own team formation rules. In most classes, students organize their own teams and quarter deadlines are set once a week. As different instructors teach this course in different class sections, we saw differences in pacing and deadlines, but these were not significant for further consideration. We invited students to participate in the research study in person or through e-mail. Interested participants were sent an email including an informed consent form and a link to the TAG. We asked students to complete the survey twice: Once at the end of Q4 and once at

the end of Q6. Our intent was to compare two measurement moments: The end of the game, when final performance outcomes are known (Q6), and the game quarter in which teams have to adapt the most (Q4), in which new technology is introduced). Unfortunately, we experienced a significant drop in response rate to the second (Q6) TAG, despite providing an additional financial incentive to complete it. As our focus is on validating the TAG, we decided to only include the Q4 responses here. A key difference between the procedure for MPL-1 and MPL-2 is that for MPL-2 we directly requested teams to communicate over Slack.

We ran the *Daedalus* game during a period of several weeks. During this period, teams can start playing at any time. We recruited participants online through local "puzzle enthusiast" communities, as well as in classrooms. Interested participants were sent a recruitment form email, which contained an informed consent form and a team registration form. Participants could start in their own full teams (having 5–7 teams members), in partial teams, or alone; we grouped interested participants in the latter two groups together so that the resulting teams had at least five members. Within the study period, teams could play *Daedalus* at their own pace. We requested that players communicate over Slack. Most teams completed the game within a week (the fastest time recorded was 4.05 hours), a few took about two weeks. Unlike *MPL*, in this study we only administered the TAG at the end of the study period due to the asynchronous nature of *Daedalus*.

5 RESULTS

5.1 Factor Analysis and Reliability

The first step in developing a new survey is determining validity and reliability. We first used a combination of confirmatory and exploratory factor analyses, a common strategy to find the best model [10, 16, 32]. Subsequently, we calculated reliability and validity measures for the resulting (sub-)scales of this model.

5.1.1 Exploratory Factor Analysis. Since the TAG was based on Rosen et al.'s [43] model (from here on called "Rosen model"), we expected that the factors from that model would be confirmed. However, the Rosen model factors were based on observations in work environments: It is possible that game environments favor different elements, which would lead to issues when confirming these factors. Moreover, the model was only used as a theoretical guide in (qualitative) observatory work, and while is useful for research contexts, the factors and their elements have not been quantitatively tested or confirmed in terms of their structure. For this reason, we first proceeded with an exploratory factor analysis (EFA) of the MPL-1 data.

We started by testing if a factor analysis was feasible. We checked for Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy for high value (close to 1.0 and not less than 0.50), conducted Bartlett's test of sphericity to test the hypothesis that the correlation matrices are an identity matrix, and looked if the data is normally distributed. The first two criteria were satisfied; however, as can be expected with ordinal data, about half of the question distributions were negatively skewed. Because of this we opted for conducting an EFA with polychoric correlations, principal axis factoring as extraction method, and varimax as rotation method.

To determine the number of factors for the EFA, we looked at a scree plot, which suggested a possible 2-factor solution, but more likely a 3-factor or 4-factor one. However, the 4-factor solution led to only two items clearly loading on a single factor; we discarded it as a feasible solution and selected the 3-factor solution. Further examination of the factor loadings (> 0.50), communalities (> 0.40), and the discrepancy between primary and secondary factor loadings ($> .20$) of the three different factor solutions led to the following conclusions. First, three items (TAGC.7, TAGC.20, and TAGC.27) did not load highly on any specific factor; as we had internally debated the value of these questions before, they were discarded. Second, the Rosen model was not clearly visible in the proposed factor model: None of our factors mirrored any of the predicted Rosen model ones. This suggests that the structure of the construct within game environments may indeed be different. Third, about half of the items load consistently high on the first factor, while the other half gets split across the other factors, suggesting strong evidence for a particular latent factor for half of the items. Based on the latter observation, we investigated theoretical explanations for these results, which did not provide a clear conclusive indication of what latent factors and structure best explain the data. For the 3-factor solution the following theoretical explanation was provided:

- **TAG Factor 1: Game Strategy.** This factor includes questions that deal with aspects of playing the game effectively, such as “making plans,” “articulating objectives,” and “identifying reasons for success and failure.”
- **TAG Factor 2: Team Planning.** This factor uses questions that deal with team interaction, teamwork, and planning.
- **TAG Factor 3: Feedback.** This factor contains questions that describe or contain the concept of feedback.

Table 1 shows what items are part of what factor in the 3-factor solution. For the 2-factor solution we came up with the following theoretical explanation:

- **TAG Factor 1: Team Gameplay.** This factor is identical to Factor 1 in the 3-factor solution.
- **TAG Factor 2: Team Organization.** This factor includes questions that involve organizing the team to play the game or reflecting on their play. This factor combines Factor 2 and 3 from the 3-factor solution.

5.1.2 Confirmatory Factor Analysis. Our next step was to conduct a series of confirmatory factor analyses (CFA) [22]. CFA is a more powerful approach than EFA in determining structure and is preferred if an *a priori* structure can be tested. For our CFA, we decided to test four theoretical models: 1-factor solution, 2-factor solution (with Team Gameplay and Team Organization), 3-factor solution (with Game Strategy, Team Planning, and Feedback), and the 4-factor Rosen model. The survey is originally based on the Rosen model, and while the EFA suggests a likely 2- or 3-factor structure, the CFA could provide conclusive evidence in favor of other models. We included a 1-factor solution option because half of the items load consistently high on a single factor, and as benchmark. The 1-, 2-, and 3-factor models are nested; the Rosen model (“R” in Table 2) is non-nested with the others, which means we can only use the fit indices for comparison with this model. We conducted CFAs with these four models, polychoric correlations, parameters freely

estimated, and WLSMV as estimator on the MPL-1, MPL-2, and Daedalus study data. We excluded a CFA with the 2- and 3-factor solution on the MPL-1 data because they have been derived from this data. Results are shown in Table 2. Due to space constraints, we report here only the model fit indices (as per [19, 22]).

The chi-squared results in Table 2 suggest that in all cases the models should be rejected ($p < .001$). However, this in itself is not sufficient to rule out a model. RMSEA results (reported with 90% confidence interval) show that the 2-factor and 3-factor solution provide a mediocre fit for MPL-2 ($< .08$), while the other models do not fit at all. All Daedalus models represent a poor fit ($< .10$) except for the 3-factor solution, which is even considered a close fit ($p = .06$). Values below .08 for RSMR are generally considered a good fit. None of the models achieve this, though the 3-factor solution consistently comes closest. Finally, CFI values above .95 are considered adequate: All of our models satisfy this index, except the 1-factor and Rosen models for MPL-2. Therefore, while the indices provide mixed results, the 2- and 3-factor solutions seem to perform better, with the 3-factor solution providing the best results.

Inspection of the coefficients and variance explained suggests that the 2- and 3-factor solutions explain for at least 20% of the variance with coefficients of at least .40, whereas other models explain less. Here too the 3-factor solution performs best overall. Note that this excepts TAGC.17, which explained very little variance in all models (less than 1% in some cases) and which was the only variable to not reach significance in the Daedalus study. Modification indices for both MPL-2 and Daedalus suggest that this item should load on TAG Factor 2 and 3 instead of TAG Factor 1. Since this item is difficult to theoretically fit into TAG Factor 1 and clearly does not fit in either of the others, we decided to remove it from the TAG.

To subsequently confirm whether the differences between the nested models are significant, we conducted scaled difference chi-squared tests [4]. Results show that for MPL-2, the 2-factor is different from both the 1-factor, $\chi(1) = 12.89$, $p < .001$, and the 3-factor, $\chi(2) = 10.17$, $p = .006$. Results are similar for Daedalus: The 2-factor is different from the 1-factor, $\chi(1) = 6.50$, $p = .01$, and the 3-factor, $\chi(2) = 10.41$, $p = .005$. Thus, the models are different, and given that the 3-factor performs overall best, we are most confident to continue with this model for our further analysis, in addition to the Rosen model. We calculated the new model without TAGC.17, which gives us the best overall results, for MPL-2: $\chi(227) = 381.78$; $p < .001$; CFI = .98; RMSEA .06 [.05-.07]; SRMR = .08, and for Daedalus: $\chi(227) = 345.30$; $p < .001$; CFI = .98; RMSEA .06 [.05-.08]; SRMR = .09. Table 1 shows the standardized loadings from these results.

5.1.3 Reliability and Validity. We calculated Cronbach’s α for the three new TAG factors (without all excluded question items) across all three studies. Table 3 shows these α -values, which are all above .7, indicating at least an acceptable internal reliability. We can only confidently say that TAG1.GS has a good reliability, a finding that is consistent with the factor analysis. We then ran Independent Samples t-tests to statistically compare these reliability values. Three statistically significant results were found: The TAG1.GS-mean for MPL-1 is significantly different from MPL-2, $t(289) = 2.57$, $p = .011$, and from Daedalus, $t(207) = 2.26$, $p = .025$; and the TAG3.FB-means for MPL-2 and Daedalus are different, $t(328) = 2.67$, $p = .007$, $df=328$). However, calculating Cohen’s d shows that none of the

Table 2: Fit indices results of the CFAs with the four different models on the three studies. F = factor solution; R = Rosen model.

Fit index	MPL-1		MPL-2				Daedalus			
	1-F	R	1-F	2-F	3-F	R	1-F	2-F	3-F	R
χ	628.70***	637.36***	1139.50***	525.11***	466.419***	995.80***	490.93***	458.72***	372.94***	452.71***
df	252	246	252	251	249	246	252	251	249	246
RMSEA	.13 [.12-.15]	.14 [.13-.15]	.13 [.12-.14]	.07 [.06-.08]	.07 [.06-.07]	.12 [.11-.13]	.09 [.08-.10]	.08 [.07-.09]	.06 [.05-.08]	.08 [.07-.09]
CFI	.95	.95	.90	.97	.98	.90	.96	.97	.98	.97
SRMR	.13	.13	.13	.09	.09	.12	.10	.10	.09	.10

*** $p < .001$ **Table 3: Comparing the three versions for reliability and validity scores. α = Cronbach's alpha; c = composite reliability; AVE = average variance extracted.**

	TAG1: GS	TAG2: TP	TAG3: FB
α - MPL-1	.94	.79	.74
α - MPL-2	.90	.76	.86
α - Daedalus	.89	.77	.81
c - MPL-1	.93	.75	.75
c - MPL-2	.88	.74	.84
c - Daedalus	.87	.77	.80
AVE - MPL-1	.57	.43	.42
AVE - MPL-2	.42	.41	.54
AVE - Daedalus	.40	.46	.46

effect sizes exceed .2 (respectively .08, .07, and .11), suggesting these differences are trivial. Consequently, we deemed the reliabilities to be consistent and satisfactorily high for this new survey.

We further considered convergent and discriminant validity through composite reliability estimates and average variance extracted (AVE), respectively (see Table 3). The composite reliability estimates (which unlike Cronbach's α considers the factor loadings) are all higher than .70 and very close to Cronbach's α values, suggesting that factor loadings do not fluctuate among items, and that we have evidence for convergent validity. In contrast, the AVE values are not all higher than (but close to) the suggested cut-off of .50 [15], implying that discriminant validity is less established.

5.2 Game Outcomes

After validating the TAG survey, the next step in our analysis was to consider the game outcomes by applying this survey to the three studies. For this analysis we considered the individual question items, the Rosen model, and the 3-factor solution. Although the factor analysis suggests the 3-factor solution, the results are not conclusive enough to fully rule out the Rosen model, on which this work was originally based. Additionally, none of our models (including the 3-factor) were strongly supported by all fit indices, which warrants further examination of the individual questions. For the individual items (reported with median and interquartile range) as well as factor items (reported with mean and standard deviation) we tested for internal consistency (data correctness and outliers) first. Below we will discuss the results from this analysis for the MPL-1, MPL-2, and Daedalus studies, respectively.

5.2.1 MPL-1. Players reported being active in MPL: Most items were either centrally distributed or skewed negatively (towards "strongly agree"). Only question TAGC.17 ($Mdn = 1$; $IQR = 1-2$) was strongly skewed positively. This question was the only negatively-written one; the positive skew suggests that emotional intra-team issues did not occur in most teams. This result also suggests that participants paid attention filling out the survey. Therefore, while we excluded this item from further consideration, it may still be considered as a control question. Another outlier is TAGC.19 ($Mdn = 3$, $IQR = 2-4$), the only question with a mode of 2 (or less than 3): This suggests that many players lacked contingency planning.

From the Rosen model's perspective, results are similar at first sight. They show that most players thought they were somewhat good at: Setting objectives and developing strategies (Plan Formulation [PF]: $M = 3.63$, $SD = 0.70$); making plans and following them (Plan Execution [PE]: $M = 3.68$, $SD = 0.69$); identifying and appraising cues (Situation Assessment [SA]: $M = 3.87$, $SD = 0.82$); and learning from each other and about the task (Team Learning [TL]: $M = 3.68$, $SD = 0.63$). However, the distributions are different. SA is more evenly distributed, suggesting that some individuals/teams were more situationally aware than others. We also see more high scores for PF than PE, suggesting that some teams were more skilled at making plans than in following through on them. These results suggest that MPL as a game supports and requires adaptation processes for all four phases.

The TAG Factors highlight starker differences. The Game Strategy (GS) scores are much higher ($M = 4.03$, $SD = 0.73$) than the scores for Team Planning (TP: $M = 3.16$, $SD = 0.85$) or Feedback (FB: $M = 3.38$, $SD = 0.69$). Because of this, the interpretation provides a more nuanced picture than the Rosen model: One where players are very engaged with thinking about how to play the game, but are less involved in planning for it and in giving feedback to each other. Inspection of the individual question items that make up the TAG Factors provide additional useful insight. For GS all items are scored similarly (all $Mdn = 4$, $IQR = 4-5$), except for TAGC.12, TAGC.16, TAGC.19 (all $Mdn = 4$, $IQR = 4-5$), which suggest that the more reflective aspects (anticipating problems, articulating strategies, contingency planning) are less engaged with by some teams. For TP (most $Mdn = 3$, $IQR = 2-4$), the most interesting observation is that TAGC.23 differs ($Mdn = 4$, $IQR = 3-4$), suggesting that of all team planning aspects, most teams did explicitly articulate responsibilities. Finally, the scores and distributions for FB are all over the place, suggesting that individuals responded differently but consistently different on these items. Particularly interesting is

that participants scored relatively high on TAGC.21 ($Mdn = 5$, $IQR = 4-5$), about whether *they* provided feedback to their teammates. This may involve a response bias or is simply a reflection of an individual bias in general as how people work in teams.

5.2.2 MPL–2. We would expect similar results for the MPL–2 study, despite a slightly different TAG version and the inclusion of Slack for team communication. Apart from the descriptive statistics, we looked here at the results from independent 2-group Mann-Whitney U tests on the individual items and the three TAG factors. From the Rosen model factors we can observe slightly higher scores for PF ($M = 3.70$, $SD = 0.60$), PE ($M = 3.78$, $SD = 0.62$), TL ($M = 3.79$, $SD = 0.56$), but especially SA ($M = 4.02$, $SD = 0.67$). The distributions look fairly similar. On the TAG factors, in contrast, we see a slight increase for GS ($M = 4.23$, $SD = 0.53$), TP ($M = 3.25$, $SD = 0.82$), but a decrease for FB ($M = 3.09$, $SD = 0.84$). However, none of these results are statistically significant, which means on a factor-level the two studies present similar results.

Consideration of the individual question items sheds more light on the differences. For GS items, it becomes apparent that two items (TAGC.3 and TAGC.10; $Mdn = 5$, $IQR = 4-5$) are agreed more strongly on, whereas now only TAGC.16 differs from the rest. This may suggest that the inclusion of Slack helped facilitate passing information and brainstorming, which is not entirely unexpected with the affordances that this platform provides. It may also suggest that it helped to reflect more. On the FB items, the responses are now more homogeneous (most $Mdn = 3$, $IQR = 2-4$), including TAGC.21. For the latter item we find a significant difference with MPL–2. We do not have an explanation for this difference. The greatest differences are to be observed with TP. Except for one item (TAGC.6), all individual items were considered statistically different, but as the scores were essentially reversed from MPL–1 overall this difference is not observed for the factor at large. This may also be a result of the “brainstorming” affordances and limitations provided by Slack, as the items more strongly agreed upon concern coming up with lists (TAGC.6 and TAGC.19) and the ones lowered are assigning tasks and responsibilities (TAGC.14 and TAGC.23).

5.2.3 Daedalus. Unlike the subtle differences between MPL–1 and MPL–2, the results for *Daedalus* are distinctly different. The Rosen factors are all lower: PF ($M = 3.01$, $SD = 0.76$), PE ($M = 3.25$, $SD = 0.56$), SA ($M = 3.38$, $SD = 0.79$), and here especially TL ($M = 2.80$, $SD = 0.72$). This suggests that playing *Daedalus* involves less plan formulation, plan execution, retrieving cues and information, and in particular less team learning. None of this is surprising, considering that many *Daedalus* puzzles involve simply doing what is asked for (“scavenger hunt”) without additional information to consider. In addition, the more asynchronous, distant teams that play this game may have provided different answers from entirely co-located teams, and the time pressure may have played a role. The TAG Factors provide a similar but even more differentiating view: GS ($M = 3.32$, $SD = 0.72$), TP ($M = 2.63$, $SD = 0.86$), but a decrease for FB ($M = 2.42$, $SD = 0.73$). What above all is striking is that both TP and FB are below 3 (“neither agree nor disagree”), which means that the majority of players disagree with being involved in team planning and giving and receiving feedback for this game. They still strategize about how to play, but less than *MPL* players.

When considering the individual question items for GS, we see a lower score for items (all $Mdn = 3$) involving resolving conflicts, anticipating problems, or improving performance (TAGC.5, 12, 15, 16, and 25)—all items we expected lower scores in, because *Daedalus* has less of a repeated structure. An interesting outlier is that players did score similarly on TAGC.3 ($Mdn = 4$, $IQR = 4-5$), which is about passing information in an efficient manner. Again, this may be a result of the use of Slack. Similarly to MPL–2, and also possibly a result of Slack, the highest-scored TP items are those items involving making lists (TAGC.6 and TAGC.19: $Mdn = 3$, $IQR = 2-4$). The lowest scores are for TAGC.14 ($Mdn = 2$, $IQR = 1-4$) and TAGC.23 ($Mdn = 2$, $IQR = 1-3$), which are about assigning tasks and responsibilities, respectively. As *Daedalus* does not require or assign roles, this result was expected and an indication that participants paid attention to filling out the survey. On the FB items, scores are indeed lower than elsewhere (most $Mdn = 2$, $IQR = 1-3$). Similar to MPL–1 and MPL–2, the items involving making use of feedback (TAGC.24 and TAGC.28: both $Mdn = 3$) score higher. Of interest is that TAGC.8, about asking for feedback, is scored relatively much lower ($Mdn = 1$, $IQR = 1-2$). This suggests that either there was less of a need to ask for feedback, or that the game’s format of asynchronous, distant teamwork created barriers to ask for feedback.

5.3 Comparative Analysis

Our last analysis compared the two different environments, *MPL* and *Daedalus*, using the TAG survey directly. The aim is to explore if the TAG survey is able to detect differences between different games, and if these differences can be explained. We used the three TAG Factors and the MPL–2 and *Daedalus* study data, considering we used the exact same survey for these studies. We looked at the distribution of responses through relative frequency histograms, displayed in Figures 1, 2, and 3 for respectively the Game Strategy (GS), Team Planning (TP), and Feedback (FB) factor. Moreover, we calculated if the differences are statistically different using independent 2-group Mann-Whitney U tests, because the distributions of the factors (in total and by game) violate assumptions of normality.

The histogram shown in Figure 1 shows that Game Strategy skews strongly towards the right (i.e., negatively skewed) for *MPL* while being more centralized for *Daedalus*, a difference confirmed statistically, $U = 3601$, $p < .001$. This factor encompasses elements such as explicitly articulating objectives (TAGC.1), making plans for the future (TAGC.5), negotiating as a team (TAGC.15), and identifying causes of failure and success (TAGC.22). All of these elements are much more prevalent in *MPL* than they are in *Daedalus*. This echoes that *MPL* requires careful planning, data interpretation, and analysis, while *Daedalus* prioritizes speed and fast throughput.

The Team Planning factor histogram shown in Figure 2 is more scattered for both games, but we see that the *MPL* score is more centralized and leaning towards the right, while *Daedalus* is slightly skewed positively. The difference is significant, $U = 7835$, $p < .001$. The distant, asynchronous design of *Daedalus* did not strongly encourage players to work together or to diversify roles as much as we had hoped; instead, we observed that the faster players would generally act as “trailblazers” solving and explaining the puzzles for the slower players. Conversely, *MPL* encouraged players to take on different roles and responsibilities and engage with different game

Figure 1: Histogram of TAG Factor: Game Strategy comparing scores from MPL (in Orange) and Daedalus (in Blue).

Figure 2: Histogram of TAG Factor: Team Planning comparing scores from MPL (in Orange) and Daedalus (in Blue).

mechanics. Even if there were no systems forcing players apart (it was possible for one player to play the game alone) we noticed that players would commonly organize themselves along “role” lines. Team Planning also incorporates elements such as preemptive planning (TAGC.11) and contingency planning (TAGC.19), which exist in both games, but are much more prevalent in *MPL*. More importantly, a likely key difference was that *MPL* players could easily meet together and plan together in person, even if they were asked to make use of *Slack* for *MPL-2*.

The Feedback factor (Figure 3) is skewed positively for *Daedalus*, but distributed for *MPL*, $U = 7152$, $p < .001$. This reinforces the finding that *Daedalus* did not really incorporate “feedback” as an element. The goal was to get the entire team to complete puzzles, meaning there was both little room and little incentive to spend time giving performance feedback. *MPL*, on the other hand, *MPL* did incentivize feedback, specifically through the quarterly performance reports, though not all teams made use of this possibility. This finding is confirmed through classroom observations and the open questions, which suggest that there were different team dynamics that may have led to varied feedback provision in games.

6 DISCUSSION

Our results suggest that TAG measures a theoretical model consisting of the factors Game Strategy (GS), Team Planning (TP), and Feedback (FB), and that it is sensitive to measure differences in games, and changes within those game environments, that may

Figure 3: Histogram of TAG Factor: Feedback comparing scores from MPL (in Orange) and Daedalus (in Blue).

affect teamwork. We will discuss these outcomes here in addition to the limitations and generalizability of this work.

6.1 Theoretical Model

We developed the TAG around the Rosen model and thus had the expectation to observe the same factors as in this model: Plan Formulation (PF), Plan Execution (PE), Situation Awareness (SA), and Team Learning (TL). However, as we were applying a business-context model built on observations to the specific context of games, we proceeded with an exploratory factor analysis (EFA) first. Indeed, the EFA and subsequent CFAs suggest that a 3-factor solution (GS, TP, FB) is more likely. Although these results are determined from three different study contexts and thus more robust than if they were to be determined from a single study context, we cannot fully rule out any one of the models tested here. We, thus, recommend that the TAG be further validated in future work.

The implications of this analysis are significant, suggesting that the theoretical model underlying these markers is different from what is proposed by Rosen et al. [43]. First, the TAG 3-factor model is *very* different from the Rosen model. PF is equally divided over GS and TP (4 items each); PE over all three (1/2 items each); SA is fully integrated into GS; and TL is equally divided over GS and FB (4 items each). Second, applying models to data gives us a different understanding. The interpretation of the *MPL-1*, *MPL-2*, and *Daedalus* study using the factor scores led to somewhat similar but still markedly different interpretations. In fact, our results suggest that the TAG model was better able to distinguish and describe the data than the Rosen et al.’s model, confirming its practical use.

A good theoretical model is not just one that adheres to the statistical criteria of fit indices, but also makes sense in terms of its constructs and what they measure, as well as its practical use in understanding the phenomenon under investigation [22]. Regarding the first, the division in GS, TL, and FB correspond well to the items measured. Future work may consider expanding the TL and FB items and make these as robust as the GS factor, which shows the most consistent results. We discuss the latter next.

6.2 Practical Applicability

We applied the TAG in three different study contexts and for two different games: *MPL* and *Daedalus*. There are four main findings from analyzing the study data that provide strong evidence for the

practical applicability of this instrument. First, the results correspond to our expectations from how the games work and what we have observed. For example, the results clearly illustrate that in *MPL* players played in roles and outlined tasks and responsibilities, whereas this was not the case in *Daedalus*. Thus, the TAG can reliably be used to measure teamwork in games.

Second, the results illustrated the improvements necessary to incentivize teamwork. If we were to redesign *Daedalus*, we would create more mechanisms to encourage “game strategy” discussions as well as “team planning” and “feedback” opportunities. We believe the game format is certainly a cause of certain low items, and would like to see more team interactions as opposed to having certain “trailblazers” be able to dominate how the game is played.

Third, the TAG can reliably distinguish teamwork in different game environments: All three TAG factors show differences between *MPL* and *Daedalus*. While it remains a question if differences between more similar games (e.g., *League of Legends* and *Dota 2*) can be detected on this factor-level, these results provide further evidence that TAG can detect how teamwork happens.

Fourth, and following up on the previous finding, on the individual item-level more subtle differences may be measurable, which is useful for not only comparing similar games but also changes within the same game. Specifically, our results suggests that the use of *Slack* may have contributed to how teamwork happened. For example, players indicated that they brainstormed and made lists much more in the *Daedalus* study than in *MPL-1*. Therefore, the TAG can be used to test new teamwork designs.

A practical application that needs further work is to observe teamwork processes, specifically adaptation, over time. We developed the TAG such that it would allow for a minimal obtrusion on the flow of the game and minimal effort on the part of players; however, in *MPL-1* we experienced a huge dropout for the second TAG request, and in *Daedalus* it was considered problematic to ask players to take the TAG more than once. To allow for repeated measures, it will possibly help to even further reduce the number of items in the future. In this work, we only focused on removing clearly problematic items.

6.3 Limitation and Generalization

Additionally, we identify three other limitations. First, one limitation is the low sample sizes. While the rule of thumb suggests it is potentially acceptable to run an EFA or CFA with as few as 50 participants, a participant amount of 100–150 is recommended, and $n = 200$ are considered ideal [22, 28]. Only the *MPL-2* study satisfies this number. However, we did not only apply the factor analyses to one sample: Taking our three study contexts together, we applied our analyses to data derived from a total of 415 participants. The possible limitation caused by the less-than-ideal sample sizes for *MPL-1* and *Daedalus* may be compensated this way. Nevertheless, applying and validating the TAG with much larger sample sizes is needed to provide more conclusive insights into its underlying theoretical model, specifically given that we could not support the initial theoretical model (i.e., Rosen et al. [43]). For future work, we recommend to continue combining EFAs and CFAs as the resulting 3-factor solution is, in fact, a new theory (see also [32]).

The second limitation, which may prevent further generalization, is the type of games studied. *Marketplace Live* (*MPL*) is a typical business simulation game: It is spreadsheet-driven and uses various performance measures to drive gameplay. *Daedalus* is more atypical: An Alternate Reality Game (*ARG*) developed with the aim to study teamwork, but grounded in familiar puzzle and scavenger hunt design. Despite sharing characteristics, neither game is very similar to well-known E-Sports games such as *League of Legends*, *FIFA*, or other team-based entertainment games. Consequently, we are unsure if the work in this study can be generalized to these kinds of games. Although future work needs to ascertain how the TAG survey thrives in other contexts, the fact that it is able to differentiate environments in an expected way gives confidence that it will be useful elsewhere too. As discussed, the survey itself can easily be adapted to other games by slightly rewording questions to the game name and context where required (see Table 1).

Finally, an inherent issue with self-report measures and Likert scales specifically are positive response biases. Additionally, there is always the question of: to what extent self-report survey responses are a reflection of actual teamwork processes. Our work shows that participants paid attention when filling out the survey, and that their responses reflect what we would expect based on the type of game and our observations of how they played. However, more systematic validation of the survey responses with respect to actual behavior is desired. This would be possible, for example, by triangulating the TAG survey responses with the behavioral game data. Another more direct opportunity is to conduct qualitative analysis on the responses on the open questions.

7 CONCLUSION

In this paper we present the TAG survey, a new instrument that measures teamwork and adaptation in games. We validated this instrument in three study contexts and found evidence for a theoretical model that can explain how teamwork takes places in games. We also demonstrated the practical use of the TAG for measuring teamwork, finding suggestions for improving teamwork in games, comparing and contrasting teamwork in different games, and evaluating new designs that aim to influence teamwork. While further validation is needed, the TAG is in a position where other researchers could make good use of it. Continued application (in different contexts) will serve to increase its strength as an instrument to measure adaptation and teamwork processes in games.

ACKNOWLEDGMENTS

We want to thank our participants and the team who built *Daedalus*. We also want to thank Gallup for helping revise the original TAG. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense, DARPA, or the U.S. Government.

REFERENCES

- [1] Aida Azadegan and Casper Harteveld. 2014. Work for or against players: On the use of collaboration engineering for collaborative games. In *Proceedings of the Workshop on Game Design Patterns, co-located with Foundations of Digital Game*. ACM.

- [2] Aida Azadegan, Casper Hartevelde, and Steven Sutherland. 2015. Design approach for collaborative decision making games. In *Proceedings of the Foundations of Digital Games*. ACM.
- [3] Shaowen Bardzell, Jeffrey Bardzell, Tyler Pace, and Kayce Reed. 2008. Blissfully productive: Grouping and cooperation in world of warcraft instance runs. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. ACM, 357–360.
- [4] Fred B Bryant and Albert Satorra. 2013. EXCEL macro file for conducting scaled difference chi-square tests via LISREL 8, LISREL 9, EQS, and Mplus. Available from the authors (2013).
- [5] C Shawn Burke, Kevin C Stagl, Eduardo Salas, Linda Pierce, and Dana Kendall. 2006. Understanding Team Adaptation: A Conceptual Analysis and Model. *Journal of Applied Psychology* 91, 6 (2006), 1189–1207. <https://doi.org/10.1037/0021-9010.91.6.1189>
- [6] ER Cadotte. 2014. Marketplace Live. Retrieved from *Innovative Learning Solutions* (www.marketplace-simulation.com) (2014).
- [7] Nancy J Cooke, Preston A Kiekel, Eduardo Salas, Renée Stout, Clint Bowers, and Janis Cannon-Bowers. 2003. Measuring team knowledge: A window to the cognitive underpinnings of team performance. *Group Dynamics: Theory, Research, and Practice* 7, 3 (2003), 179.
- [8] Nancy J Cooke, Eduardo Salas, Janis A Cannon-Bowers, and Renee J Stout. 2000. Measuring team knowledge. *Human Factors* 42, 1 (2000), 151–173.
- [9] Bart A De Jong, Kurt T Dirks, and Nicole Gillespie. 2016. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of Applied Psychology* 101, 8 (2016), 1134.
- [10] Jamie DeCoster. 1998. Overview of factor analysis. (1998). Retrieved June 26th 2019 from <http://www.stat-help.com/notes.html>.
- [11] Anders Drachen, Matthew Yancey, John Maguire, Derrek Chu, Iris Yuhui Wang, Tobias Mahlmann, Matthias Schubert, and Diego Klabajan. 2014. Skill-based differences in spatio-temporal team behaviour in defence of the ancients 2 (dota 2). In *2014 IEEE Games Media Entertainment*. IEEE, 1–8.
- [12] Francis T Durso, Carla A Hackworth, Todd R Truitt, Jerry Crutchfield, Danko Nikolic, and Carol A Manning. 1998. Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly* 6, 1 (1998), 1–20.
- [13] Mica R Endsley. 1988. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting*, Vol. 32. SAGE Publications Sage CA: Los Angeles, CA, 97–101.
- [14] Joel R Evans and Anil Mathur. 2005. The value of online surveys. *Internet research* 15, 2 (2005), 195–219.
- [15] Claes Fornell and David F Larcker. 1981. Structural equation models with unobservable variables and measurement error: Algebra and statistics.
- [16] David W Gerbing and Janet G Hamilton. 1996. Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal* 3, 1 (1996), 62–72.
- [17] Ruthanna Gordon. 2015. *Alternate Reality Games For Behavioral and Social Science Research*. ETC Press.
- [18] Jamie C Gorman, Nancy J Cooke, Jennifer L Winner, Jasmine L Duran, Harry K Pedersen, and Amanda R Taylor. 2007. Knowledge training versus process training: The effects of training protocol on team coordination and performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 51. SAGE Publications Sage CA: Los Angeles, CA, 382–386.
- [19] Daire Hooper, Joseph Coughlan, and Michael Mullen. 2008. Structural equation modelling: Guidelines for determining model fit. *Articles* (2008), 2.
- [20] Zaheer Hussain and Mark D Griffiths. 2009. The attitudes, feelings, and experiences of online gamers: a qualitative analysis. *CyberPsychology & Behavior* 12, 6 (2009), 747–753.
- [21] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 2017. What Makes a Strong Team? Using Collective Intelligence to Predict Team Performance in League of Legends. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work*. 2316–2329.
- [22] Rex B Kline. 2015. *Principles and practice of structural equation modeling*. Guilford publications.
- [23] Yubo Kou and Xinning Gui. 2014. Playing with strangers: understanding temporary teams in league of legends. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-Human Interaction in Play*. ACM, 161–169.
- [24] Steve WJ Kozlowski, James A Grand, Samantha K Baard, and Marina Pearce. 2015. Teams, teamwork, and team effectiveness: Implications for human systems integration. (2015).
- [25] Field Day Lab. 2018. ARIS: Create location-based games and stories. <https://fielddaylab.org/make/aris/>
- [26] Whitney E Lewis. 2018. Design Scaffolding for Computational Making in the Visual Programming Tool ARIS. (2018).
- [27] Timothy C Lisk, Ugur T Kaplanali, and Ronald E Riggio. 2012. Leadership in multiplayer online gaming environments. *Simulation & Gaming* 43, 1 (2012), 133–149.
- [28] John C Loehlin. 1998. *Latent variable models: An introduction to factor, path, and structural analysis*. Lawrence Erlbaum Associates Publishers.
- [29] Peter V Marsden and James D Wright. 2010. *Handbook of survey research*. Emerald Group Publishing.
- [30] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. 2000. The influence of shared mental models on team process and performance. *Journal of Applied Psychology* 85, 2 (2000), 273.
- [31] John E Mathieu, Scott I Tannenbaum, Jamie S Donsbach, and George M Alliger. 2014. A review and integration of team composition models: Moving toward a dynamic and temporal framework. *Journal of Management* 40, 1 (2014), 130–160.
- [32] Masaki Matsunaga. 2010. How to factor-analyze your data right: do's, don'ts, and how-to's. *International Journal of Psychological Research* 3, 1 (2010), 97–110.
- [33] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.
- [34] Robert J Mislevy, Andreas Oranje, Malcolm I Bauer, Alina von Davier, Jiangang Hao, Seth Corrigan, and Erin Hoffman. 2014. *Psychometric Considerations in Game-based Assessment*. GlassLab.
- [35] Bonnie Nardi and Justin Harris. 2006. Strangers and friends: Collaborative play in World of Warcraft. In *Proceedings of the 2006 20th anniversary Conference on Computer Supported Cooperative Work*. ACM, 149–158.
- [36] Fernando Felix do Nascimento Junior, Allan Sales da Costa Melo, Igor Barbosa da Costa, and Leandro Balby Marinho. 2017. Profiling Successful Team Behaviors in League of Legends. In *Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web*. ACM, 261–268.
- [37] Robert E Ployhart and Paul D Bliese. 2006. Individual adaptability (I-ADAPT) theory: Conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability. In *Understanding adaptability: A prerequisite for effective performance within complex environments*. Emerald Group Publishing Limited, 3–39.
- [38] Nataliia Pobiedina, Julia Neidhardt, Maria del Carmen Calatrava Moreno, and Hannes Werthner. 2013. Ranking factors of team success. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 1185–1194.
- [39] Nataliia Pobiedina, Julia Neidhardt, Maria del Carmen Calatrava Moreno, Laszlo Grad-Gyenge, and Hannes Werthner. 2013. On successful team formation: Statistical analysis of a multiplayer online game. In *Business Informatics (CBI), 2013 IEEE 15th Conference on*. IEEE, 55–62.
- [40] Patrick Prax. 2010. Leadership style in World of Warcraft raid guilds. *Proceedings of DiGRA Nordic* (2010).
- [41] Elaine D Pulakos, Sharon Arad, Michelle A Donovan, and Kevin E Plamondon. 2000. Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology* 85, 4 (2000), 612.
- [42] Daniel Rapaport. 2017. What to expect from the booming esports industry in 2017. *Sports Illustrated* (2017). <https://www.si.com/tech-media/2017/02/09/esports-industry-expectations-billion-dollar>
- [43] Michael A Rosen, Wendy L Bedwell, Jessica L Wildman, Barbara A Fritzsche, Eduardo Salas, and C Shawn Burke. 2011. Managing adaptive performance in teams: Guiding principles and behavioral markers for measurement. *Human resource management review* 21, 2 (2011), 107–122.
- [44] Eduardo Salas, Terry L Dickinson, Sharolyn A Converse, and Scott I Tannenbaum. 1992. Toward an understanding of team performance and training. (1992).
- [45] Eduardo Salas, Carolyn Prince, David P Baker, and Lisa Shrestha. 1995. Situation awareness in team performance: Implications for measurement and training. *Human factors* 37, 1 (1995), 123–136.
- [46] Matthias Schubert, Anders Drachen, and Tobias Mahlmann. 2016. Esports Analytics Through Encounter Detection Other Sports. (2016).
- [47] Magy Seif El-Nasr, Bardia Aghabeigi, David Milam, Mona Erfani, Beth Lameman, Hamid Maygoli, and Sang Mah. 2010. Understanding and evaluating cooperative games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 253–262.
- [48] Kimberly A Smith-Jentsch, Janis A Cannon-Bowers, Scott I Tannenbaum, and Eduardo Salas. 2008. Guided team self-correction: Impacts on team mental models, processes, and effectiveness. *Small Group Research* 39, 3 (2008), 303–327.
- [49] Renée J Stout, Janis A Cannon-Bowers, Eduardo Salas, and Dana M Milanovich. 1999. Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors* 41, 1 (1999), 61–71.
- [50] Zachary O Toups, Jessica Hammer, William A Hamilton, Ahmad Jarrah, William Graves, and Oliver Garretson. 2014. A framework for cooperative communication game mechanics from grounded theory. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-Human Interaction in Play*. ACM, 257–266.
- [51] Harald Warmelink. 2014. *Online gaming and playful organization*. Routledge.
- [52] David K Woods and Paul G Dempster. 2011. Tales from the bleeding edge: The qualitative analysis of complex video data using Transana. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, Vol. 12.
- [53] Jason Wuertz, Scott Bateman, and Anthony Tang. 2017. Why Players Use Pings and Annotations in Dota 2. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1978–1988.

- [54] Pu Yang, Brent E Harrison, and David L Roberts. 2014. Identifying patterns in combat that are predictive of success in MOBA games. In *Proceedings of the Foundations of Digital Games*.
- [55] Mark C Zorzie. 2012. *Individual adaptability: Testing a model of its development and outcomes*. Michigan State University. Psychology.