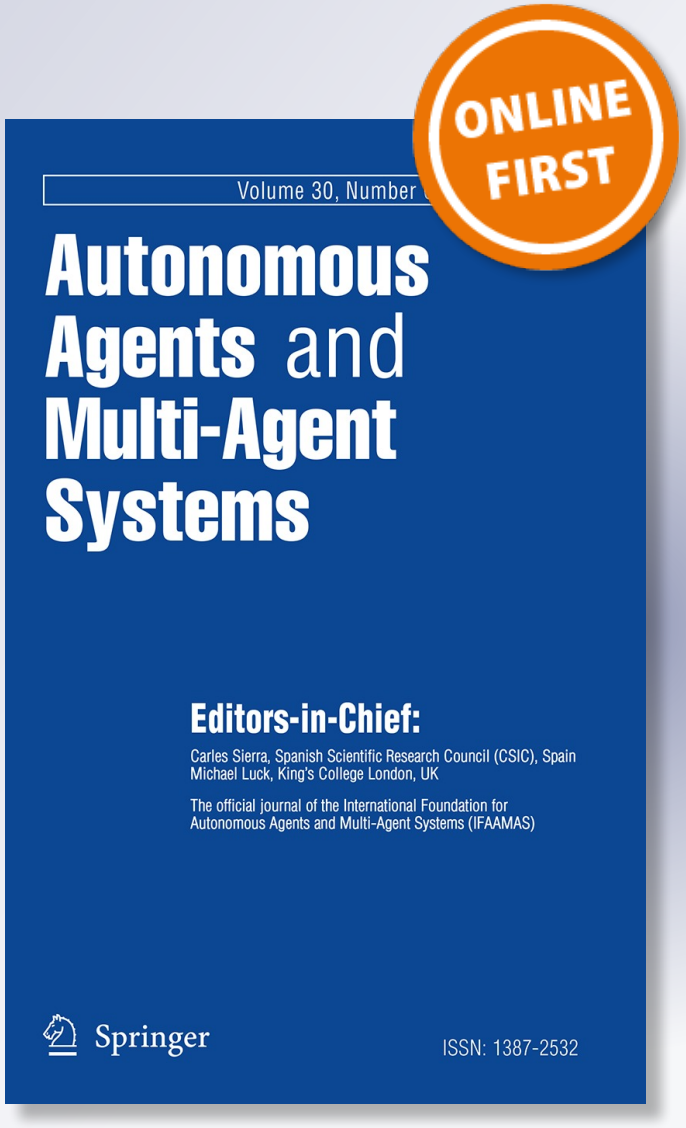
Three years of the RoboCup standard platform league drop-in player competition

Katie Genter, Tim Laue & Peter Stone

Autonomous Agents and Multi-Agent Systems

ISSN 1387-2532

Auton Agent Multi-Agent Syst
DOI 10.1007/s10458-016-9353-5



Springer

Auton Agent Multi-Agent Syst
DOI 10.1007/s10458-016-9353-5

Three years of the RoboCup standard platform league drop-in player competition

Creating and maintaining a large scale ad hoc teamwork robotics competition

Katie Genter[1] · **Tim Laue**[2] · **Peter Stone**[1]

Abstract The Standard Platform League is one of the main competitions at the annual RoboCup world championships. In this competition, teams of five humanoid robots play soccer against each other. In 2013, the league began a new competition which serves as a testbed for cooperation without pre-coordination: the Drop-in Player Competition. Instead of homogeneous robot teams that are each programmed by the same people and hence implicitly pre-coordinated, this competition features ad hoc teams, i.e. teams that consist of robots originating from different RoboCup teams and as such running different software. In this article, we provide an overview of this competition, including its motivation, rules, and how these rules have changed across three iterations of the competition. We then present and analyze the strategies utilized by various drop-in players as well as the results of the first three competitions before suggesting improvements for future competitive evaluations of ad hoc teamwork. To the best of our knowledge, these three competitions are the largest annual ad hoc teamwork robotic experiment to date. Across three years, the competition has seen 56 entries from 30 different organizations and consisted of 510 min of game time that resulted in approximately 85 robot hours.

Keywords Ad hoc teamwork · Coordination · Multiagent teamwork · RoboCup · Robot soccer

✉ Katie Genter
katie@cs.utexas.edu

Tim Laue
tlaue@uni-bremen.de

Peter Stone
pstone@cs.utexas.edu

[1] Department of Computer Science, The University of Texas at Austin, Austin, TX 78712, USA

[2] Department of Computer Science, University of Bremen, 28359 Bremen, Germany

Published online: 19 November 2016

1 Introduction

As robots become more prevalent in the world, they are increasingly being designed to work in teams to accomplish tasks. One such example is delivery robots utilized in hospitals, such as the Aethon TUG robot.[1] Another example is the Amazon Robotics Kiva robots that move products to and from box packers in warehouses [25]. Usually, all of the robots on a team are programmed by one organization, and hence are implicitly designed to work together. RoboCup, an annual international robotics competition, features many such teams that are programmed by universities, companies and other organizations to play soccer in various leagues [8]. This article presents a specific competition held in the RoboCup Standard Platform League (SPL), namely the Drop-in Player Competition.

In the Drop-in Player Competition, each team programs a robot to coordinate with unknown teammates. The teams are asked to not pre-coordinate, so that during games these agents have to engage in *ad hoc teamwork* in order to reason about their teammates' abilities and intentions in real time and determine how to best assist their team. Each agent's goal should be to win the soccer game while being judged as a 'good teammate' by human observers.

It is often challenging when working with teams of real robots to gather extensive experimental data. Over three years, the SPL Drop-in Player Competition has seen 56 entries from 30 organizations, involved at least 50 human participants, and consisted of 38 games for a total playing time of 510 min. With 10 robots scheduled to participate in each game, this totals to an experiment utilizing roughly 85 robot hours. Hence, this series of Drop-in Player Competitions proved to be the largest ad hoc teamwork experiment on robots that the authors are aware of to date, and is likely one of the largest robotic experiments involving 30 different organizations across three years.

The SPL Drop-in Player Competition grew from a technical challenge held at RoboCup 2013 in three different leagues [10]. The 2013 SPL technical challenge was optional for teams participating in the SPL, and hence only saw six SPL teams participate. Furthermore, the 2013 challenge was announced with little advance notice so many teams did not have time to tailor their strategies to the ad hoc setting. The authors of this article helped plan, organize, and run the 2013 technical challenge as well as the substantially larger SPL Drop-in Player Competitions at RoboCup 2014 and RoboCup 2015. Both of these larger SPL Drop-in Competitions were mandatory for teams participating in the SPL and announced well in advance. This article details all three SPL Drop-in Player Competitions, highlights the advances in each year of the competition, and discusses various drop-in player strategies utilized in the competition.

This article makes two major contributions by (1) presenting the SPL Drop-in Player Competition's setup, rules, and scoring metrics across three iterations of the competition and (2) summarizing and analyzing the participating teams' strategies and comparing their performance in multiple Drop-in Player Competitions with their performance in multiple main competitions. The article is laid out as follows. Section 2 describes the SPL as a RoboCup league and introduces the concept of ad hoc teamwork. Details pertinent to the Drop-in Player Competition are discussed in Sect. 3 while the scoring schemes utilized for each competition are presented in Sect. 4. The strategies employed by various drop-in players in each competition are described in Sect. 5. Section 6 presents the results of the 2013, 2014, and 2015 competitions and analyzes these results. Section 7 suggests improvements for

[1] http://www.aethon.com/.

subsequent ad hoc teamwork competitions. Section 8 situates this research in the literature and Sect. 9 concludes the article.

2 Background

Two important areas of background knowledge are introduced in this section. The first is the Standard Platform League (SPL) of RoboCup and the second is the multiagent systems research area of ad hoc teamwork.

2.1 RoboCup standard platform league (SPL)

RoboCup is an annual international robotics competition that had its first competition in 1997 [8]. RoboCup has many leagues, ranging from RoboCup @Home which involves single robots acting intelligently in a home environment to multiple simulation leagues to various physical robot soccer leagues. Almost all of the leagues involve autonomous agents that must act in their environment without any human interaction.

The Standard Platform League (SPL) is different from other RoboCup leagues in that all teams must use the same robotic platform. This effectively makes the SPL a software competition, albeit on a robotic platform. Hence, although teams must implement their software on real robots, they are required to use particular versions of the SoftBank NAO[2] and they are not allowed to physically alter the NAO in any manner.

Historically, the SPL has allowed teams to compete in a main team competition as well as various technical challenges. The main team competition is usually executed as one or more round robins where the top teams from these round robin pools gain spots in an 8-team single-elimination bracket. The technical challenges are optional competitions lasting no more than two hours each in which teams compete in announced challenge tasks that are designed to advance the league. A small Drop-in Player Competition was held as a technical challenge in 2013 before becoming a separate SPL competition in 2014 and 2015.

Teams in the SPL compete in 5 on 5 soccer games on a 9 meter by 6 meter soccer field, as depicted in Fig. 1. Each game consists of two 10-min halves. Teams must play completely autonomously—no human input is allowed during games outside of game state signals sent by an official to communicate to the robots when a goal has been scored, when they have been penalized, etc. The playing environment is partially color-coded and the robots on each team are allowed to communicate with each other over a wireless network.

2.2 Ad hoc teamwork

Since 1997, RoboCup has served as an excellent domain for testing teamwork, coordination, and cooperation. Most teams have successfully programmed their robots to work well as a team, coordinating which robot should go to the ball, which robot should play defense, and even what formation should be adopted against various opponent types. However, the 2013 drop-in player challenge [10] was one of the first organized efforts to evaluate a player's ability to coordinate with a set of teammates in an ad hoc manner, and the 2014 and 2015

[2] https://www.ald.softbankrobotics.com/en.

Fig. 1 NAO robots playing in an SPL game during RoboCup 2014

SPL Drop-in Player Competitions greatly improved upon the 2013 challenge in both scale and participation.

Ad hoc teamwork's defining characteristic is its focus on creating agents that can cooperate with unknown teammates without prior coordination. Stone et al. imagined 'staging field tests of ad hoc team agents at the annual RoboCup competitions' in their 2010 AAAI challenge paper that introduced ad hoc teamwork [21]. The SPL Drop-in Player Competitions at RoboCup 2013, RoboCup 2014 and RoboCup 2015 did just this. By organizing the SPL Drop-in Player Competition as a full-fledged competition, the authors and the RoboCup organization as a whole have created the potential for a long-standing empirical testbed for ad hoc teamwork research.

In addition to being a good testbed for ad hoc teamwork research, development of players with strong ad hoc teamwork capabilities may assist the RoboCup community in reaching its ultimate goal. Specifically, the ultimate goal of RoboCup is to have a team of fully autonomous humanoid robot soccer players win a FIFA regulation soccer game against the most recent World Cup champions by 2050 [7]. As robots become larger and more expensive, and the field grows larger, most organizations and universities will likely not have the space, money, or manpower to field an entire team of these robots. Hence, as RoboCup nears its 2050 goal, increasingly more RoboCup teams may actually be ad hoc teams formed of drop-in players from multiple universities with limited pre-coordination. The limited pre-coordination may be a result of unwillingness or impracticality to use a single, shared codebase, different approaches for robot decision making, difficulty in communication, or other issues. In addition, these teams might be formed on quite short notice because of the malfunction of some players. As such, the SPL Drop-in Player Competition can also be seen as contributing towards the ultimate RoboCup goal.

3 Competition description

One of the goals of the SPL Drop-in Player Competition is to adapt the SPL to serve as the largest testbed and source of data for ad hoc teamwork using real robots. Although the Drop-in Player Competition is based on the main RoboCup SPL soccer competition, several changes were necessary to make it a meaningful competition about teamwork without pre-coordination. In this section, we discuss these changes as well as the standard communication scheme and the organization of the competition.

3.1 Altered rules of the game

For the most part, the rules of the Drop-in Player Competition games [12,14,16] are the same as for main competition games in the SPL. In fact, the only major difference concerns role assignment. In normal SPL games, there is a designated goalkeeper robot on each team. Such a predefined role assignment assigns a particular player to be the goalkeeper instead of forcing the players to arrange the role assignments for themselves.

However, in the Drop-in Player Competition the players need to arrange the role assignments for themselves. Hence, the first robot that enters its own penalty area is considered the goalkeeper for the remainder of the game. This does not tie this robot to acting as goalkeeper—nor prevent any other robot from attempting to play goalkeeper—but instead allows the referee to afford the goalkeeper specific privileges given to just the goalkeeper. For example, the goalkeeper is able to push robots while defending the ball and is also able to play the ball with its hands while in the penalty box. Note that this rule may lead to games in which no goalkeeper exists since its existence is not enforced externally. However, this strategy is not unheard of in the SPL—team HTWK has chosen strategies in the past where they favored an additional field player over a goalkeeper. Forcing the players to arrange their own role assignments can also lead to games in which multiple robots believe they are the goalkeeper. In 2013 this rule would result in the 'goalkeeper' that did not enter the penalty area first being repeatedly penalized and removed from the field as an 'illegal defender'. However, in 2014 and 2015 this robot would be allowed to remain on the field as the 'illegal defender' rules were changed such that multiple robots were permitted in the goal box simultaneously.

3.2 Standard communication

The SPL introduced progressively advanced wireless standard communication interfaces for use in each iteration of the Drop-in Player Competition. Use of this standard communication interface was optional in 2013. In 2014, the standard communication interface was declared to be mandatory, but there was no system utilized to check for compliance. In 2015, a system was implemented to ensure that drop-in players that did not send messages according to the standard communication interface were not allowed to play until they were seen to be sending messages that included and updated the required fields. A short description of this system is given in [19] and the open source code is available in the SPL's infrastructure repository on GitHub (Fig. 2).[3]

Each message was required to follow the predefined format specified by the standard communication interface for the competition. Requiring a particular format enables team-oriented planning by serving as a communication interface. The message fields for each

[3] https://github.com/bhuman/GameController.

Fig. 2 Team Communication Monitor during a normal Standard Platform League game. The centered 3-D view displays the current world states, i.e. position, status, and ball position, communicated by the playing robots. The side views show the status and the frequency of each robot's communication

competition are provided in Table 1—if similar fields have different names or slight changes in definition across the competitions, the most recent name or definition is listed. The required standard communication interfaces for 2014 and 2015 are available in the SPL's infrastructure repository on GitHub[4] while the interface for 2013 is available on the SPL website.[5]

3.3 Organization of the competition

In each drop-in game, ideally all 10 robots on the field originate from different teams. Although this ideal was obtained in the 2014 and 2015 competitions, it was impossible to obtain in the 2013 competition because only 6 teams entered the competition. To handle this situation, drop-in teams in the 2013 competition contained at most two players originating from the same team.

To achieve scores that reliably reflect the drop-in capabilities of a single robot, it is best to play as many games as possible with as many different teammates and opponents as possible. To substantiate this intuitive statement, organizers of the 2014 drop-in challenge in the RoboCup 3D Soccer Simulation League ran experiments to determine just how many games were needed. The RoboCup 3D Soccer Simulation league uses simulated NAO robots to play 11 versus 11 games. Since it is relatively easy to run many simulated games in the 3D Soccer Simulation league, organizers were able to empirically determine how many games were needed to achieve statistically meaningful goal differences. They found that it took roughly half the total number of possible permutations of drop-in player team pairings before their goal difference results stabilized.[6] In the SPL, we do not have the ability to play nearly that number of games, so we utilize judge scoring in addition to goal difference when determining the best drop-in players.

The team assignments were randomly generated by hand in 2013 and by an algorithm that was also used for the 3D Soccer Simulation League drop-in games in 2014 and 2015

[4] https://github.com/bhuman/GameController/blob/master/include/SPLStandardMessage.h.

[5] https://www.tzi.de/spl/pub/Website/Downloads/pickup.h.

[6] MacAlpine, P. (2014). Private communication.

Table 1 Required message fields for each competition

Field	Meaning	2013	2014	2015
PlayerNum	Assigned player number (1–5)	Yes	Yes	Yes
TeamColor	Color of team (red, blue)	Yes	Yes	No
TeamNum	Assigned team number (98,99)	No	No	Yes
Fallen	1 if robot is fallen, 0 otherwise	Yes	Yes	Yes
Penalized	Seconds the robot has been penalized, −1 otherwise	Yes	No	No
Pose [3]	Pose and orientation of robot	Yes	Yes	Yes
PoseVariance [3]	Variance in the robot's pose	Yes	No	No
WalkingTo [2]	Robot's target position on field	No	Yes	Yes
ShootingTo [2]	Target position of next shot	No	Yes	Yes
BallAge	Seconds since this robot last saw the ball	Yes	Yes	Yes
Ball [2]	Position of ball relative to robot	Yes	Yes	Yes
BallVel [2]	Velocity of the ball	Yes	Yes	Yes
Suggestion [5]	Suggestions for teammate roles	No	No	Yes
Intention	What role the robot intends	No	Yes	Yes
AverageWalkSpeed	The robot's average speed	No	No	Yes
MaxKickDistance	The robot's maximum kick distance	No	No	Yes
CurrentPositionConfidence	The robot's current self-location confidence	No	No	Yes
CurrentSideConfidence	The robot's current directional confidence	No	No	Yes

(Algorithm 1 in [10]). Each player was not assigned to play with every other player, as this would require many more games than we could run at the competition. For example, in 2014 27 games would have been required in order for each player to play with every other player at least once. Table 2 shows the relevant characteristics of our schedule at each of the three Drop-in Player Competitions discussed in this article.

In addition to generating games with robots from up to 10 different teams, each game requires four referees as well as 3-6 judges (depending on the competition year). Judges and referees are always selected from teams that are not playing in a match if possible, and it was preferred that judges all originate from different teams. Hence, running a single Drop-in Competition game would involve participants from 9 to 18 different RoboCup teams (again, depending on the competition year)!

To avoid any pre-coordination, the assignment of drop-in players to teams was announced as close to each match time as possible. At the start of the tournament, only the time slots of the matches were announced to allow teams to prepare for these time slots. The time of the announcement regarding which robots would play on which teams in a match varied between 30 min and multiple hours prior to the match. This range depended on each day's overall schedule and the need to inform all participants in enough time to avoid any misunderstandings. In addition to announcing the assignments of players to teams as late as possible, all participants were also explicitly told to refrain from pre-coordinating. We have no reason to suspect that any participants attempted to pre-coordinate.

Table 2 Statistics from each competition

	RoboCup 2013	RoboCup 2014	RoboCup 2015
Games	4	15	19
Teams	6	25	27
Entries/team	6–7	6	7–8
Min teammates	4	18	21
Max teammates	5	22	26
Min opponents	5	17	19
Max opponents	5	20	24

'Games' refers to the number of total games held in the Drop-in Player Competition, 'Teams' refers to the number of drop-in players participating, and 'Entries/team' refers to the number of games scheduled for each unique drop-in player. 'Min Teammates' denotes the minimum number of teammates any drop-in player has, 'Max Teammates' denotes the maximum number of teammates any drop-in player has, and likewise for 'Min Opponents' and 'Max Opponents'

4 Scoring scheme

Agents designed for the Drop-in Player Competition should be adept at reasoning about their teammates' abilities and intentions and responding in such a way that helps their team the most. The Drop-in Competition scoring metrics discussed in this section were carefully designed to reward agents for being good teammates and not just for having better low-level skills. However, even with thoughtful planning, designing fair scoring metrics was difficult.

Scoring metrics were one aspect of the Drop-in Player Competition that we iteratively improved between each competition. Hence, although the scoring scheme was different for each competition, there were also many similarities. One such similarity is the use of human judges. Human judges were utilized to help identify good teamwork abilities in agents in order to ameliorate the problems of random variance in the games and the quality of teammates affecting the goal differences.

In the following sections, we first present the scoring schemes utilized in all three Drop-in Player Competitions and then we discuss the three scoring metrics utilized in these competitions in detail.

4.1 2013 Scoring scheme

The 2013 Drop-in Player challenge was scored using two metrics:

- average goal difference
- average human judged score

For each game, three judges were asked to award each drop-in player a teamwork score ranging from 0 to 10, where 10 means the drop-in player was an excellent teammate and 0 means that the drop-in player ignored its teammates. The judges were instructed to focus on teamwork capabilities, rather than individual skills, such that a less-skilled robot could still be given top marks for good teamwork. Unlike in subsequent competitions, the same three judges observed all four 2013 Drop-in Player Competition games.

The average goal difference was normalized and then added to the average judge score in order to determine the overall winner of the challenge. The normalization factor was computed to be

$$\frac{10}{\text{best average goal difference of a drop-in player}}$$

Each drop-in player's average goal difference was then multiplied by this normalization factor in order to get each drop-in player's normalized average goal difference.

Full details related to the 2013 scoring scheme can be found in the Technical Challenges rulebook [12].

4.2 2014 Scoring scheme

The 2014 Drop-in Player Competition was scored using the same two high-level metrics as in 2013: average goal difference and average human judged score. However, the way the judge score was calculated and obtained was different.

For this competition, six judges observed each game. Three judges observed each team for a half. At halftime, the team each judge was scoring changed to reduce noise. This procedure resulted in six scores from six different judges per robot per game. Whereas judges in 2013 attempted to judge all 10 players on the field at once, judges in 2014 were asked to just watch one team for each half in this competition. This change was made so that the judges could better focus their attention in order to judge fairly and consistently.

The judges were asked to score the players based on the following criteria with the goal that providing criteria would help lead to consistent judging across various judges:

- Judged constantly during the game:
  - Appropriate decision to pass to teammate = +1 to +4
  - Receiving a pass = +1 to +3
  - Pushing a teammate = −2 (whether this occurs must be determined by each judge)
  - Unclassified bonus or penalty = −2 to +2 (capped at −10/+10 per half; requires justification by judge)
- Judged once per half:
  - Game participation = −10 to +10 (requires justification by judge)

Passing was specifically rewarded because it is a major manifestation of soccer cooperation. Note that a range of possible points is specified, which allows the judges to award the significance and brilliance of a pass. In any case, judges were told that passes that do not provide any benefit for the team should not be rewarded. Receiving a pass is an expression of useful positioning.

Pushing a teammate, which is judged using the same criterion as player pushing in the main competition, is penalized because well positioned players should rarely be in a position from which pushing a teammate is possible.

Additionally, each judge may also assign each robot an additional score between −2 and +2 at any point for actions that are beneficial or harmful for the team but are not covered by the other rules. Judges were explicitly notified that players should not be rewarded (or punished) for scoring a goal, as doing so is rewarded (or punished) via goal differential.

At the end of each half, judges were able to assign each robot a reward or punishment between −10 and +10 for its participation in that half. Judges were told that good positioning or actively contributing to the game should be rewarded, while poor positioning, harmful behaviors, or inactivity should be punished.

Finally, for any half in which a drop-in player did not leave the sideline, judges were asked to give that robot a score of −20 for that half.

After all drop-in games are complete and the average goal difference and average judge score have been computed for each drop-in player, the two scoring metrics are normalized and added up to determine the overall winner. Specifically, the normalization occurs as follows:

- The player with the highest average goal difference receives 100 points and the player with the lowest average goal difference receives 0 points.
- The player with the highest average human judged score receives 100 points and the player with the lowest average human judged score receives 0 points.
- All other average goal differences and average human judged scores are scaled linearly.
- Each player's judge and goal points are added.

Full details related to the 2014 scoring scheme can be found in the 2014 rulebook [14].

4.3 2015 Scoring scheme

The 2015 Drop-in Player Competition was scored using two metrics:

- average game result
- average human judged score

The average game result was calculated by awarding points at the end of each game based on the result: Win, Draw, Loss, Absent. Players on the winning team received 2 points, players on both teams received 1 point if the game result was a draw, and players on the losing team received 0 points. Players who were assigned to be on one of the playing team, but who did not enter the field during the game, received -2 points.

This average game result metric replaced the average goal difference metric used in 2013 and 2014 to put (1) more emphasis on wins and losses and (2) less emphasis on games with lopsided results.

In an attempt to use fewer judges and improve consistency, we used five judges for each Drop-in Player Competition game in 2015. The most senior judge served as head judge. The head judge (1) instructed the judges in how to properly judge the game, (2) monitored the drop-in player communication monitor introduced in Sect. 3.2 to ensure that robots were not on the field if they were not communicating, and (3) made notes throughout the game regarding when players were not present on the field.

The four remaining judges evaluated each player on the field. The judges scored the players based on the following criteria, which were designed to be simpler than the criteria used in 2014. If a robot was on the field for at least 5 min of the 10 min half, then its judge score minimum was capped at 0. If a robot was not on the field at all for the half, its overall judge score was automatically -5.

- Judged constantly during game:
  - Positive team play = +1 to +4
    - Examples: passing, positioning to receive a pass or intercept an opponent's shot
  - Negative team play = -1 to -4
    - Examples: pushing teammates, taking the ball away from teammates
- Judged once per game:
  - Overall positioning (poor 0, average 5, exceptional 10)
  - Game participation (mostly inactive/not on field 0, average 5, exceptional 10)

For overall positioning and game participation, the default score for any robot should be 5. Only in cases of remarkably negative or positive behavior should scores of 0 or 10

be awarded. These instances of remarkable behavior are what the judges should watch for though.

The two scoring metrics are normalized and added up to determine the overall winner of this competition as follows:

- The team that has the highest average game result will get 100 points.
- The team that has the lowest non-negative average game result will get 0 points.
- All other non-negative average game results become scaled linearly.
- The team that has the highest average human judged score will get 200 points.
- The team that has the lowest non-negative average human judged score will get 0 points.
- All other non-negative average human judged scores become scaled linearly.
- Each team's judge and game result points will be added.
- Teams with negative average game results or negative average judge score will be excluded from the final rankings. This constraint has been added as the 2014 competition showed that teams with an excessively negative score in one of the two scoring metrics have a too strong influence on the overall scoring due to the linear scaling (cf. Sect. 6.2).

Full details related to the 2015 scoring scheme can be found in the 2015 rulebook [16].

4.4 Goal difference scoring

In the 2013 and 2014 Drop-in Player Competitions, each drop-in player's average goal difference was calculated as the average goal difference of the games in which the player was scheduled to compete. In this section we discuss the importance of goal difference in drop-in player scoring and discuss some difficulties experienced with goal difference scoring.

4.4.1 Importance

Goal difference is a useful scoring metric because it embodies the main aspect of being a good teammate—helping your team win by as much as possible.

4.4.2 Difficulties

One of the main difficulties that affected average goal differences was that not all drop-in players who registered for the Drop-in Competition showed up. Due to extremely late notice by the missing drop-in players, their spots remained empty which resulted in not all games being played 5 versus 5. Although this did likely affect the goal difference in these games, we believe the effect was not necessarily significant because their absences were spread across various teammates and opponents.

One of the main difficulties with using a player's average goal difference for determining the best drop-in player is that all players on a team receive the same goal difference from a game despite some players impacting the final game result more than others. In the case of the SPL Drop-in Competition, even players who did not enter a game received credit for the game in terms of goal difference if they were scheduled to play in it. The problem with this is highlighted by the existence of players who missed many games and yet still received better-than-expected goal difference ranks.

Finally, average goal differences were sometimes skewed by one or two lopsided games. To counteract this, we considered average game result in 2015, which is explained below.

4.5 Game result scoring

In the 2015 Drop-in Player Competition, each player's average game result was calculated to be the average result—win, draw, or loss—across all games in which the player was scheduled to compete. In this section, we discuss the importance of game results in drop-in player scoring as well as discuss some related difficulties.

4.5.1 Importance

Game result scoring measures what actually matters: did the team our player joined win, lose, or draw?

4.5.2 Difficulties

Game result scoring suffers from many of the same difficulties as goal differential scoring. One additional difficulty that game result scoring suffers from is additional clustering of scores. This clustering can cause many players to receive the same game result normalized score. Although this is usually not troublesome in itself, it means that the game result scores do not differentiate between players as much as goal difference scoring.

4.6 Judge scoring

Human judges were utilized in each of the three SPL Drop-in Player Competitions discussed in this article. Judges were instructed to watch and score every game with the intention of evaluating the teamwork abilities of some subset of the players. In this section, we discuss why human judges were utilized in this competition as well as difficulties that we experienced.

4.6.1 Importance

The Drop-in Player Competition is about creating good teammates. Hence, players should be rewarded for good teamwork and not just superior low-level skills. Despite having a standard platform in the SPL, some participants have designed superior walk engines and kick engines that could give them an advantage if only goal difference or game result were considered. Hence, human judges were used to recognize good teamwork that might otherwise be overlooked.

In general, we found that although there was a correlation between the goal difference score or game result score and the judge score, the correlation was not always strong and hence the judge score represented a distinct quality.

In the 2013 Drop-in Player competition, the judge ranks and goal difference ranks had a Pearson correlation coefficient (R) of 0.8286, meaning they were strongly positively correlated. A strong positive correlation means that a player with a particular goal difference rank was likely to have a similar judge rank. However, the judge ranks and goal difference ranks for the 2014 Drop-in Player Competition had a R of 0.3618, meaning they were weakly positively correlated. Lack of a strong correlation implies that judge scoring represents a different quality than goal difference scoring. Finally, in 2015, the correlation coefficient increased to 0.5120 (considering only the 24 teams that achieved scores greater than 0).

Despite the strong correlation in 2013 and the moderate correlation in 2015, judge scoring does still provide the ability to easily recognize teams displaying good teamwork. As such,

although judge scoring has difficulties, which we discuss next, it is an important part of the SPL Drop-in Player Competition.

4.6.2 Difficulties

It is difficult to design and enforce a scoring scheme for human judging that (1) fairly assesses teamwork capabilities, (2) is usable for human judges, and (3) can be consistently applied across various judges.

Although some of the scores 'required' the human judges to provide justification for their scores in 2014, judges very rarely provided justification. From personal experience and conversations with judges at the competition, this was likely due to both the speed of the game and the fact that most judges wanted to finish judging as quickly as possible. Although justifications would have allowed the judging criteria to be improved for future competitions, there is no feasible way in which to require judges to give justifications. As such, judges were not asked to provide justifications in 2015.

During each competition, teams were assigned to provide judges for certain matches. Some teams notified their judge shortly before each game, leaving the judge very little time to become familiar with the judging criteria. Although having the same judges at each match provides more consistency, and with just four games in 2013 was feasible, judging duties in 2014 and 2015 were distributed across all of the participating teams in order to not place an undue time burden on any particular individuals or teams. As a result, judges were often confused about when to award bonuses and penalties to individual robots. Additionally, human judging is inherently subjective and inconsistent. As a result, despite averaging across multiple judges for each game, the judge scores were likely only a loose approximation of each robot's ability as a teammate.

In Sect. 7.3 we discuss some options for improving the quality and consistency of human judging as well as an option that could reduce the need for human judging.

5 Drop-in player strategies

One of the most interesting aspects of ad hoc teamwork and the Drop-in Player Competition is observing how different agents attempt to contribute to the team they join. After the 2013 competition was held on a small scale, one of the large questions that remained was: *What strategies did various drop-in players employ?*

In order to answer this question in subsequent competitions, each participating team was asked to submit a short description of the strategy they used in the competition after the 2014 and 2015 competitions. These strategies were then publicly released on the SPL website [13,15] in order for teams to learn from each other, prepare for subsequent competitions, and provide other researchers a better overview of the current status of the competition.

5.1 2014 Player strategies

Immediately after the 2014 Drop-In Player Competition concluded, every participating team was asked to submit a short one-paragraph description of the strategy they used in the competition. In total, 17 out of 23 participants submitted a description. In this section we discuss some interesting communication, coordination, and behavior trends gathered from analyzing these strategies.

5.1.1 Communication and coordination

As described in Sect. 3.2, all robots within a team are connected by a wireless network and are able to send standardized messages to each other. In theory, the content of these messages should be a valuable source of information when coordinating with teammates. However, in practice, proper communication may not be established because:

- not all robots actually send messages
- not all robots fill in all of the standard message elements
- some robots send incorrect data, likely as a result of mis-localization, false positive ball observations, or improper use of the standard message elements

In their strategy description, more than half of the participants do not mention these problems or explicitly state that they trust their teammates. However, eight participants mentioned that they do not accept all communicated messages:

- Berlin United, HTWK, and HULKs state that they discard *most* of the information that they receive but they do not discuss how they determine *which* information to discard.
- MiPal did not implement the communication interface.
- Nao Devils and Philosopher send messages but discard all incoming messages.
- B-Human and Northern Bites implemented approaches to determine the reliability of their teammates by checking the plausibility of the transmitted information and the teammates' ability to fulfill their announced roles, respectively. However, Northern Bites did not use this implementation during the competition.

As described in the next section, this limited communication affected the chosen strategies in multiple cases. Communication also seemed to have an impact on the success of players, as discussed in Sects. 6.2.2 and 7.2.

5.1.2 Typical player behaviors

There appears to be one strategy applied by the majority of the drop-in players: *Play the ball if it is close and/or no other robot wants to play the ball. Take a supporting position otherwise.* In many cases, the decision to go to the ball depends on the communicated positions and intentions of teammates. The chosen supporting positions vary from simple strategies like *Stay close to the ball* to more complex computations involving teammate positions. These strategies are, as mentioned by multiple participants, often the same strategies used for their main competition games.

However, some of the participants that accepted limited or no messages from their teammates used a special strategy to avoid conflicts, and thus possible negative scores, with teammates that want to play the ball. They positioned their robots at fixed positions on the field, e.g. a defensive position inside their own half or somewhere close to the field's center, and waited for the ball to appear nearby. If it did, the robot would attempt to kick or dribble the ball towards the opponent goal. Otherwise, the robot would remain at its position and track the ball.

One role that was only mentioned in a few descriptions, and rarely seen in games, was the goalkeeper. Some participants claimed they actively avoided this role because they believed that the scoring scheme disadvantaged goalkeepers by limiting their judge scoring potential.

We discuss the behaviors and strategies of the winning players in Sect. 6.2.2. In Sect. 7.2 we present various areas for future behavior improvements.

5.2 2015 Player strategies

Similarly to the previous year, every team participating in the 2015 Drop-In Player Competition was asked to submit a short one-paragraph description of the strategy they used in the competition. In total, 23 out of 27 participants submitted a description. In this section we discuss the trends found upon analyzing these submitted strategies.

5.2.1 Communication and coordination

As described in Sect. 3.2, from 2015 on, all robots within a team are required to send standardized messages to each other. Robots that do not send a minimum amount of messages (at least 20 messages every minute) or do not fill in the required fields are not allowed to play. These new constraints have been added to overcome some of the problems described in Sect. 5.1.1. Nevertheless, it cannot be required that the transmitted information is correct or has a certain precision. Therefore, several robots send information but choose to use a limited amount of the information they receive.

In their strategy description, 19 participants describe their usage of the communicated messages:

- Seven participants do not use any information that they receive. Only one of them (UNSW Australia) explicitly mentions that they do not believe their teammates.
- Nine participants mention that they use information provided by their teammates.
- Austrian Kangaroos, B-Human, and Nao Devils mention the usage of approaches to determine the reliability of their teammates by checking the plausibility of the transmitted information.

Overall, the communication among the robots appears to have increased since 2014. As was seen in 2014, communication continued to have an impact on the success of players in 2015—we discuss this further in Sects. 6.3.2 and 7.2.

5.2.2 Typical player behaviors

The 2015 behavior strategies of drop-in players can be divided into four classes:

- Seven participants use their *normal player behavior*, often including minor workarounds to compensate for missing teammate information and pre-coordination. Six of these participants are among those who use all or filtered information from teammates; the seventh participant did not provide any information about communication. These robots are able to play all roles that a normal player would be able to play.
- Eight participants have implemented a *special behavior* for the Drop-in Player Competition, each including the ability to play different roles such as striker or supporter.
- Five participants take the simple behavior of trying to *find the ball and kick* it towards the opponent goal, disregarding any teammates. Unsurprisingly, four of them are among those who do not use any information from teammates; the fifth participant did not provide any information about communication.
- Three participants focus on *defensive positioning*. Their robots always stay inside the own half and kick the ball towards the opponent goal if it appears in their proximity.

Out of these main strategies, implementing a special behavior for the Drop-in Player Competition generally led to better performance of the team. Although adopting a defensive positioning strategy would theoretically lead to better team performance, unfortunately most

of the players that attempted to use this strategy were not effective defenders. Overall, in contrast to 2014, one can observe an increased trust in the capabilities of the teammates and therefore an increase in the complexity of the applied behaviors.

Probably as a result of the changed scoring scheme, the acceptance of the goalkeeper role seems to have increased and more robots were seen choosing to actively take this role. In their strategy descriptions, seven participants explicitly mention the goalkeeper. Only one of them (B-Human) deliberately avoids becoming goalkeeper, as they assume they are able to contribute more to the team as a field player. The other six teams take this role for a variety of reasons. Interestingly, some of the strategy descriptions that describe rather simple strategies strongly imply avoidance of the goalkeeper role.

We discuss the behaviors and strategies of the winning players in Sect. 6.3.2. In Sect. 7.2 we consider potentially fruitful areas for future behavior improvements.

5.3 The reliability of teammates

When trying to cooperate with other robots, it is crucial to have as much information about their current state and intentions as possible. This is why the Standard Platform League introduced the standard communication interface, which all robots were encouraged to use since 2014 and required to use since 2015.

Despite encouraging and then requiring usage of the standard communication interface, most teams' strategies are still based on a general mistrust of the information received from teammates. This is probably because many robots are known to have self-localization problems—especially after introducing the white goals in 2015—and thus send information that is not precise enough or is even incorrect. Although there is other information in the standard packet, most elements depend on the knowledge of one's own position, such as the derived global position of the ball as well as the positions a robot intends to walk to or shoot to.

As it is not allowed to preconfigure a robot with information about which other robots are expected to be capable of sending trustworthy information, a player needs to derive this information during the course of play. One approach for this is to compare one's own observations with the received information from other players. For instance, players from the B-Human team compare their own ball observations with the ball observations of each teammate [18]. If multiple matches occur within a certain amount of time, the respective teammate is considered trustworthy and upcoming actions are chosen with regard to the position and intention of that teammate. Ball observations are not the only information that could be evaluated though. Players could also compare robot observations with the transmitted robot positions. However, as all robots look the same, identification is difficult. In theory, it could be possible to detect the numbers on the robot jerseys, but so far, no team has applied such an approach.

A few teams decide to coordinate not based on their trust of a teammate, but on the location of the teammate. Specifically, some teams only cooperate with robots that they observe near the ball. Unfortunately, this was revealed in personal communication and was not documented in the short strategy descriptions. As such, there is no information regarding how these teams actually coordinate with the robot near the ball.

One final approach that could be used to coordinate without using communication sent by other players would be to observe the entire field visually and keep track of all teammates. However, given the current robot platform, this would be a very difficult task. The NAO's field of view is quite limited, covering only a small fraction of the field. Additionally, the

Table 3 Scores for the 2013 SPL Drop-in Player Competition (listed from best ranked to worst)

Team	Country	Judge avg	Goal diff avg	Goal diff norm	Drop-in comp score	Main comp rank
B-Human	Germany	6.67	1.17	10	16.67	1
Nao Devils	Germany	6.24	0.57	4.87	11.11	5–8
rUNSWift	Australia	5.22	0.67	5.73	10.95	4
UT Austin Villa	USA	6.00	−0.29	−2.48	3.52	3
UPennalizers	USA	4.48	−0.57	−4.87	−0.39	17–22
Berlin United	Germany	3.38	−1.29	−11.03	−7.65	9–16

NAO's low camera resolution makes it difficult to accurately identify other robots over longer distances.

6 Results and analysis

The previous section considered the various strategies utilized by drop-in players across the last three competitions. In this section, we consider the competition results and determine how well these strategies fared. In particular, we consider the effect of different scoring schemes across the three competitions as well as whether a team's success in the main team competition was highly correlated with the team's success in the Drop-in Player Competition.

Throughout this section, some similar columns will appear in the results tables (Tables 3, 4, 5) for each of the three competitions. Two columns refer to judge scores: *Judge Avg* columns give the raw average scores given by judges while *Judge Norm* columns give the normalized *Judge Avg*, where the normalization is done as described in Sect. 4. Two columns refer to goal difference scores: *Goal Diff Avg* columns give the average goal differences across the games in which the player was scheduled to play, while *Goal Diff Norm* columns give the normalized *Goal Diff Avg*, where the normalization is done as described in Sect. 4. The 2015 competition uses the notion of average game result and hence Table 5 features two columns that refer to average game result scores: *Game Result Avg* and *Game Result Norm*. The *Game Result Avg* column gives the average game result across the games in which the player is scheduled to play, while the *Game Result Norm* column gives the normalized *Game Result Avg*, where the normalization is done as described in Sect. 4.3. Finally, each table also includes a *Drop-in Comp Score* column which depicts the player's overall Drop-in Player Competition score and a *Main Comp Rank* column which depicts the player's originating team's ranking in the main team competition (if they participated in the main team competition).

Now that the columns for all of the results tables have been introduced, in the following sections we consider the results from each Drop-in Player Competition as well as analyze these results.

6.1 2013 Results and analysis

The 2013 SPL Drop-in Player Competition only included six teams. The results of this competition were calculated as detailed in Sect. 4.1 and are displayed in Table 3. This section analyzes these results.

Table 4 Scores for the 2014 SPL Drop-in Player Competition (listed from best ranked to worst)

Team	Country	Judge avg	Judge norm	Goal diff avg	Goal diff norm	Drop-in comp score	Main comp rank
B-Human	Germany	4.72	100	1.33	100	200	3
HTWK	Germany	1.28	83.04	1.00	89.47	172.51	2
Nao Devils	Germany	1.61	84.68	0.67	78.95	163.63	5–8
TJArk	China	2.17	87.41	0.50	73.68	161.10	9–12
Berlin United	Germany	−0.58	73.87	0.67	78.95	152.82	5–8
DAInamite	Germany	0.08	77.15	0.50	73.68	150.84	13–20
UPennalizers	USA	0.67	80.03	0.33	68.42	148.45	9–12
Austrian Kangaroos	Austria	−2.90	62.45	0.83	84.21	146.66	9–12
rUNSWift	Australia	3.00	91.52	−0.17	52.63	144.15	1
Cerberus	Turkey	0.72	80.30	0.00	57.89	138.20	13–20
Northern Bites	USA	−1.81	67.85	0.33	68.42	136.27	13–20
NTU RoboPAL	Taiwan	1.61	84.68	−0.50	42.11	126.78	5–8
UT Austin Villa	USA	−1.28	70.45	−0.17	52.63	123.08	13–20
HULKs	Germany	−1.83	67.72	−0.17	52.63	120.35	13–20
UnBeatables	Brazil	−3.36	60.19	0.00	57.89	118.09	–
RoboCanes	USA	−1.06	71.55	−0.50	42.11	113.65	13–20
Philosopher	Estonia	−0.25	75.51	−0.67	36.84	112.36	13–20
Edinferno	UK	−0.08	76.33	−0.83	31.58	107.91	13–20
MiPal	Australia/Spain	−0.94	72.09	−1.00	26.32	98.41	–
SPQR	Italy	−8.00	37.35	0.00	57.89	95.24	9–12
MRL	Iran	−1.22	70.73	−1.33	15.79	86.52	5–8
UChile	Chile	−4.50	54.58	−1.83	0.00	54.58	4
UTH-CAR	Mexico	−15.6	0.00	−0.50	42.11	42.11	–

As can be seen in Table 3, although the intention was for the judge score and the goal difference score to be weighted equally, the goal difference was more heavily weighted than expected. This occurred because (1) the judge averages were used directly instead of being normalized and (2) the goal difference averages were normalized poorly. Note that the judge average scores—which were used directly when calculating the overall score—all fall in the range of 3.38 to 6.67. These scores were then added to the goal difference normalized scores, which has a range of −11.03 to 10. Due to this large difference in ranges, the goal difference scores had a much larger effect on the overall Drop-in Competition scores than the judge scores.

6.1.1 Comparison of competition rankings

With only six teams participating in the Drop-in Player Competition, it is difficult to say anything conclusive about the relative performance of the drop-in players. However, we did find a moderate positive correlation between how teams performed in the main team

Table 5 Scores for the 2015 SPL Drop-in Player Competition (listed from best ranked to worst)

Team	Country	Judge avg	Judge norm	Game result avg	Game result norm	Drop-in comp score	Main comp rank
HTWK	Germany	15.0	197.5	1.71	100.0	297.5	3
B-Human	Germany	15.2	200.0	0.86	50.0	250.0	2
Nao Devils	Germany	10.4	137.2	1.29	75.0	212.2	9-12
Berlin United	Germany	11.3	148.6	1.00	58.3	207.0	5-8
NTU RoboPAL	Taiwan	7.4	98.0	1.43	83.3	181.4	5-8
HULKs	Germany	3.9	52.0	1.57	91.7	143.7	13-20
UChile	Chile	7.0	92.2	0.71	41.7	133.9	4
Northern Bites	USA	6.1	80.6	0.86	50.0	130.6	13-20
UNSW	Australia	9.0	118.3	0.14	8.3	126.6	1
Austrian Kangaroos	Austria	5.6	74.0	0.86	50.0	124.0	13-20
Philosopher	Estonia	6.1	80.3	0.71	41.7	121.9	13-20
RoboEireann	Ireland	5.7	75.4	0.71	41.7	117.1	13-20
SPQR	Italy	6.1	81.1	0.29	16.7	97.7	13-20
TJArk	China	5.3	69.3	0.43	25.0	94.3	9-12
UT Austin Villa	USA	6.4	85.0	−0.14	0.0	85.0	5-8
UnBeatables	Brazil	6.3	83.3	−0.13	0.0	83.3	–
MRL	Iran	4.8	62.7	0.29	16.7	79.4	9-12
Cerberus	Turkey	5.0	66.6	0.14	8.3	74.9	5-8
Edinferno	UK	3.9	51.4	0.29	16.7	68.0	–
WrightOcean	China	4.4	58.4	0.14	8.3	66.8	–
Z-Knipsers	Switzerland	4.3	57.3	0.00	0.0	57.3	–
JoiTech	Japan	2.9	38.2	0.14	8.3	46.5	–
RoboCanes	USA	3.4	44.6	−0.14	0.0	44.6	9-12
UPennalizers	USA	3.3	43.8	−0.43	0.0	43.8	13-20
Camellia Dragons	China	2.2	0.0	−1.14	0.0	0.0	–
Blue Spider	China	−5.0	0.0	−2.00	0.0	0.0	–
Linkoping	Sweden	−5.0	0.0	−2.00	0.0	0.0	13-20

competition and how they performed in the Drop-in Player competition; the correlation coefficient between the ranks is 0.7302. This means that there is a tendency for a team who does well in the main competition to do well in the Drop-in Player Competition. Notably, B-Human finished first in both competitions while UPennalizers and Berlin United finished in the bottom half of both competitions. However, with just six teams participating, the correlation coefficient may not be representative of a larger scale Drop-in Player Competition.

6.2 2014 Results and analysis

The results of the 2014 SPL Drop-in Player Competition were calculated as detailed in Sect. 4.2 and are displayed in Table 4. This section analyzes these results in detail.

One of the goals of the SPL Drop-in Player Competition is for a team comprised of the top five drop-in players to be able to play comparably to the winner of the main SPL competition.

At RoboCup 2014 we held the first of these 'All Star' games where robots from B-Human, HTWK, Nao Devils, TJArk and Berlin United played as a drop-in team against the 2014 SPL main competition champion rUNSWift in a full-length game. The result was 4-2 in favor of rUNSWift, but the relative closeness of the result shows that the drop-in team did relatively well. In the main competition, rUNSWift allowed only one goal while scoring 42 goals across 7 games. rUNSWift's closest game in the main competition was a 5-1 win against HTWK in the championship game. Hence, the fact that the drop-in team was able to take rUNSWift to a 4-2 result is indeed impressive.

When looking at the *Judge Avg* column in Table 4, the fact that UTH-CAR had a substantially worse *Judge Avg* than any other drop-in player stands out. UTH-CAR had such a low judge score because their robot was inactive or not on the field for most of its games and hence was rated as a poor teammate. Drop-in players that do not appear on the pitch for a half automatically receive a −20 judge score for that half (−15.58 implies that UTH-CAR did not show up for most halves). Other teams, such as UChile and SPQR, also failed to put an active robot on the field for some of their games, and hence received rather poor judge scores. UTH-CAR's substantially lower *Judge Avg* had a large impact on the results of the competition because it caused 22 of 23 drop-in players to have a *Judge Norm* of greater than 54 and 17 of 23 drop-in players to have a *Judge Norm* of greater than 70. This caused the judge scores to have a weaker influence on the overall Drop-in Player Competition results than desired.

The top three teams with regard to *Judge Avg* were B-Human with an average score of 4.72 per half, rUNSWift with an average score of 3.00 per half, and TJArk with an average score of 2.17 per half. Their submitted strategy descriptions [13] note that all three teams decided to either play the ball or take a supporting role based on their own perceptions and the information communicated by teammates. These are simple, yet effective, Drop-in Player Competition strategies.

6.2.1 Comparison of competition rankings

We can compare each team's Drop-in Player Competition rank (teams are listed from best ranked to worst in Table 4) to their main SPL competition rank (*Main Comp Rank* in Table 4). These two ranks have a Pearson correlation coefficient (R) of 0.3021, meaning that they are weakly positively correlated. Three teams did not compete in the main competition, and hence do not have rankings for the main competition.

In general, better teams in the main competition did tend to perform better in the Drop-in Player Competition—only one team that finished tied for 13th in the main competition was in the top 9 teams in the Drop-in Player Competition. Interestingly though, some teams who performed very well in the main competition, namely MRL and UChile, finished in the bottom three teams for the Drop-in Player Competition. This suggests that solid low-level skills but deployment of normal game code will not necessarily yield success in the Drop-in Player Competition.

6.2.2 Analysis of the winning players

When inspecting the final ranking of the competition, the accumulation of German teams at the top ranks attracts attention: the top three drop-in players (B-Human, HTWK, and Nao Devils) as well as the 5th and 6th placed drop-in players (Berlin United and DAInamite) come from Germany, with TJArk from China at the 4th place as the only exception.

Some of these teams, namely B-Human, Nao Devils, and TJArk, have based their software on the same framework published by the B-Human team [17]. However, this framework does not provide any particular components that might benefit the participation in drop-in games. In addition, many other teams, which performed significantly worse in this competition, also use this framework. Thus, the success of the top teams does not seem to be related to any particular software framework.

One commonality among four of the top five teams (B-Human, HTWK, Nao Devils, and Berlin United) is their participation in the Drop-In Competition at the RoboCup German Open 2014. This competition was the largest Drop-in Competition test run under realistic conditions prior to RoboCup 2014—the scenes shown in the supplementary video[7] were recorded at this competition. Eight teams played four drop-in games in which a preliminary version of the 2014 rule set was applied. Furthermore, three of these teams (B-Human, Nao Devils, and Berlin United) participated in the 2013 technical challenge version of the Drop-in Player Competition. In this early state of the Drop-In Competition, one can assume that the experiences gained by these teams in these competitions impacted their performance in the 2014 Drop-in Player Competition.

Another noteworthy aspect is the fact that three of the top teams expressed a very limited trust in the communicated messages of others (HTWK and Berlin United) or tried to estimate the reliability of their teammates (B-Human), as described in Sect. 5.1.1. This strategy appears to have been one of the insights gained at previous Drop-in Player Competitions: *avoid cooperation with unreliable teammates*. As mentioned in Sect. 5.1.1, in the 2014 competition it could not be assumed that all robots comply with the standard communication interface. Indeed, there are likely multiple teams with wrong or incomplete implementations that enable them to communicate well with their normal teammates, but poorly with drop-in teammates.

Finally, it should be noted that there was a large variance in the amount of additional code teams implemented for the Drop-in Player Competition. In some cases, developing the team's drop-in player was a team member's main contribution. However, the strategy descriptions [13] summarized in Sect. 5.2.2 also show that some teams just slightly altered their normal game code.

6.3 2015 Results and analysis

The results of the 2015 SPL Drop-in Player Competition were calculated as discussed in Sect. 4.3 and are displayed in Table 5. This section discusses these results in detail.

As in 2014, an 'All Star' game between the top 5 Drop-in Player competition players and the main team competition champion was also held in 2015. The members of the 2015 'All Star' team included HTWK, B-Human, Nao Devils, Berlin United and NTU RoboPAL. The result of the game was a 5:0 win for UNSW Australia, who was referred to as rUNSWift in previous competitions. However, more concerning than the score was the fact that the 'All Star' team appeared to play rather poorly together. Indeed, multiple players would attempt to play the ball at the same time, and no players seemed to be playing offense or defense consistently away from the ball.

As can be seen in Table 5, the scoring metric was designed such that the judge score was weighted to be twice as influential as the game result when calculating the final score. However, due to the normalization process of each of these, the judge scores ended up being more heavily weighted than expected. This can particularly be seen by the fact that (1) B-

[7] http://www.informatik.uni-bremen.de/agebv2/downloads/videos/genter_laue_stone_iros_15.mp4.

Human finished in 2nd place despite having a 50% worse normalized game result score that HTWK and (2) HULKs had the second best game result, yet finished 6th overall.

6.3.1 Comparison of competition rankings

Similarly to 2014, 2015 saw a limited correlation between how teams performed in the main team competition and how they performed in the Drop-in Player competition. Considering all teams that participated in both competitions and scored in the Drop-In Competition, the correlation coefficient between the ranks is 0.3578. There are a few specific data points to mention. First, consider main competition champion UNSW Australia—despite winning the main team competition, they finished 9th in the Drop-in Player Competition. Their game result was the part of the scoring metric where they did poorly, meaning that the teams they were part of generally did poorly. Second, consider Nao Devils. Despite not finishing in the top 8 teams in the main team competition, they finished 3rd in the Drop-in Player Competition.

6.3.2 Analysis of the winning players

Most—4 out of 5—of the 'All Stars' in 2015 were 'All Stars' in 2014 as well, namely HTWK, B-Human, Nao Devils, and Berlin United. This would imply that these teams have strategies and behaviors that are well-suited for the Drop-in Player competition. This might also imply that these teams consistently put the necessary time and manpower towards the Drop-in Player Competition. Moreover the same four teams also participated in the Drop-in Competition that was held during the RoboCup German Open 2015, a few months prior to RoboCup 2015, where they held ranks 1–4 out of 8 participating teams. Experience has shown that testing under realistic conditions contributes to a better adaptation to the scenario. In a way, this may explain why the 'All Stars' played well in normal drop-in games but failed to play well together in the 'All Star' game. Perhaps these teams had been programmed to be wary of teammates and were unable to adequately adapt when given more capable teammates.

The top four teams are among the participants that rely on communication with their teammates. Two of these teams (B-Human and Nao Devils) also state that they filter incoming information from teammates. This is an improvement from 2014, as neither Nao Devils nor Berlin United relied on communication in previous competitions. All four teams used non-trivial behaviors that allowed their robots to switch between different roles. B-Human and Berlin United applied their normal game strategies, while Nao Devils and HTWK appeared to utilize specialized implementations for the 2015 Drop-In Competition. Compared to 2014, B-Human did not change its strategy significantly. However, the other three top teams seemed to apply more complex strategies in 2015 than they had in the previous competition. In the cases of Nao Devils and Berlin United, this is probably a result of beginning to use information communicated by teammates.

Surprisingly, the 5th place team—NTU RoboPAL—was among the participants that did not use any communicated information and applied only a simple strategy that consisted of finding and kicking the ball.

7 Lessons learned

In running and observing three years of the SPL Drop-in Player Competition, we can provide insights regarding (1) how to set up a similar competition, (2) strategy improvements

that would likely be beneficial to teams competing in such competitions, (3) improvements organizers can make in subsequent competitions, and (4) how experiences from the Drop-in Player Competition can apply to general ad hoc teamwork research. Although our experience is from organizing the SPL Drop-in Player Competition, most of the insights discussed in this section apply to any competitive ad hoc teamwork evaluation.

7.1 Organizing a similar competition

Organizing a competition, including designing the rules and scoring scheme, can be very difficult and time-consuming. In this section, we provide some lessons we learned while organizing the first three years of the SPL Drop-in Player Competition as well as some suggestions for those organizing similar competitions. These suggestions would be most valuable to other RoboCup leagues looking to begin a Drop-in Player Competition, but could apply to other types of competitions as well.

Running the drop-in player technical challenge in three leagues [10] in 2013 before starting the much larger SPL Drop-in Player Competition in 2014 had multiple benefits. First, it allowed us to introduce the idea of drop-in games to RoboCup in a manner that garnered support. Second, the organizers of three leagues were able to work together to design the general competition design. Third, it allowed us to evaluate the feasibility of the SPL competition in a small scale, low risk, 2-h technical challenge.

After running the SPL Drop-in Player Competition as an optional technical challenge in 2013, the league organizers required all SPL teams to participate in the Drop-in Player Competition in 2014 and 2015. Teams were encouraged to do well in the Drop-in Player Competition because their performance was loosely tied to qualification decisions the following year—doing very well would guarantee teams a spot in the main team competition the following year while doing exceptionally poorly would take away pre-qualification for the following year earned via any other means. By requiring all SPL teams to participate in 2014 and 2015, the competition became large enough and the play advanced enough for multiple teams to consider their drop-in player behavior as a research contribution separate from their normal team behavior. We do not believe the Drop-in Player Competition would be as successful and popular if it has not been mandatory for all SPL teams in 2014 and 2015. The competition has advanced enough that almost all teams applied to participate in the—no longer mandatory—Drop-in Player Competition at RoboCup 2016.

Requiring teams to report their strategies publicly has allowed the competition to advance rapidly because teams can review all of the strategies from previous competitions while designing their own strategy for subsequent competitions. Somewhat surprisingly, most teams submitted strategies when asked to do so despite submission not being mandatory for most teams. We believe we achieved such a high strategy submission rate because (1) we only asked teams to write a single paragraph and (2) we asked them to do so within a 24-hour window of the final Drop-in Player Competition game. However, it might be beneficial for the competition if the top teams were required to publish their strategies through a formally archived proceeding, as this would provide other researchers insight into the strategies employed and the reasons these strategies were employed.

We found that checking whether drop-in players were sending valid and complete messages was necessary to reach compliance—merely telling teams their player must comply (as was done in 2014) did not achieve compliance. Hence, if there is a critical component of the competition that is not controlled by the organizers but can be checked, it is likely worth the effort to design tools (such as our Team Communication Monitor described in Sect. 3.2) to check this component automatically.

As mentioned in Sect. 5, one characteristic that dominated team strategies was mistrust in the information received from other players. When organizing a similar competition, one should definitely pay attention to this issue. If the competition is set in a scenario in which all agents can be assumed to have a good knowledge about their state within the global frame of reference, e.g. because they receive position information from an external trustworthy source or because state estimation is significantly easier than in the RoboCup Standard Platform League, this problem might be less fundamental but may shift towards 'Do my teammates come to the *right* conclusions?' Otherwise, one should make sure that the environment as well as the communication scheme contain observable elements that allow the agents to infer whether information from others appears to be trustworthy or not.

The most difficult part of designing the SPL Drop-in Player Competition was defining the scoring scheme. We updated the scoring scheme between each main RoboCup competition, but we also often updated the scoring scheme after the German Open and/or the US Open (both of which are held a few months before each main RoboCup competition). Since it was infeasible to run drop-in games outside of competitions, and doing so at competitions required agreement and effort from multiple teams, the sample size for evaluating any particular scoring scheme was incredibly small. As such, the lessons to be learned are (1) attempt to simulate your competition if possible and determine how well your scoring metrics, normalization, and weighting work together given various results and (2) take full advantage of any opportunities to test your scoring metrics at preliminary competitions (such as the German Open and US Open).

Not all ad hoc teamwork evaluations require human judges. Ideally, a scoring scheme could be designed that is fair and consistent without requiring human judges. However, in some cases, such as in the SPL Drop-in Player Competition where only a limited number of games can be scheduled and many participants have different baseline abilities, human judges become necessary. Hence, when a scoring scheme utilizes multiple scoring metrics, it is important to normalize each metric and then weight each metric as desired. In the 2015 SPL Drop-in Competition we decided to weight the judge score twice as much as the game result score, as we believed judges could better identify good drop-in players. Hence, in competitions with high variance in the baseline behaviors of participants, we recommend highly weighting judge scores because judges can easily identify good teamwork even in weaker participants.

7.2 Suggested strategies

The previous section discussed strategies for organizing a competitive ad hoc teamwork evaluation similar to the SPL Drop-in Competition. In this section, we turn our attention to participants in competitive ad hoc teamwork evaluations and provide some suggestions for improving their strategies.

7.2.1 Determining whether to trust teammates

As we discussed in Sect. 5.2.1, seven participants did not use communicated information from their teammates in the 2015 competition. One team explained that they did not trust the information sent by their teammates—hence, we would recommend that participants send complete and accurate information to their teammates as much as possible. The other six participants did not explain why they did not use information from their teammates, so we would recommend that these participants attempt to validate and/or utilize this information.

A few participants considered the trustworthiness of each teammate's communication in the 2014 and 2015 SPL Drop-in Player Competitions, and in general these participants did well. Hence, both results and intuition suggest that evaluating the trustworthiness of each teammate would likely be very useful. So far, most participants that do evaluate trustworthiness seem to either accept or reject all communicated information from a teammate, but it would likely be more useful to instead consider information from less-trustworthy teammates at a lesser confidence level instead of completely disregarding it. Further research and development of all teams regarding self-localization and object tracking—which are not major foci of the Drop-in Player Competition—might lead to a situation in which the problem of reliability is less important because all players are inherently more reliable. Additionally, most major potential changes of the competition environment that impact state estimation—such as white goals (2015), a realistically looking ball (2016), and a voluntary outdoor competition (2016)—have already been implemented. As it is expected that future competition updates will not focus on aspects that make state estimation more difficult, one might expect a saturation and convergence of solutions in this area within the next few years.

Certainly there are many interesting related research questions. Approaches that perform compound team state estimation under high uncertainty are worth investigating. In this scenario, high uncertainty does not only mean a degree of noise but also a significant number of false positives as well as false negatives. Furthermore, robust approaches for behavior selection given a high amount of uncertainty and partial observability of the environment are a topic on which research efforts should be spent. Finally, although a few teams have begun to consider how to determine the trustworthiness of teammates, most of these approaches are relatively simplistic. There are plentiful research opportunities to consider substantially more communicated data and learn about the abilities of teammates online when determining how much to trust particular teammates. Additionally, determining how to use this data when determining how to act within a team is an open research question.

7.2.2 Seeking teamwork opportunities

Very few drop-in players attempted to pass to teammates in the 2013, 2014, and 2015 SPL Drop-in Player Competitions, but a few attempted passes were witnessed at the 2015 competition. As such, in subsequent competitions players may want to seek out situations in which it would be reasonable to make or receive a pass as this could help players receive better judge scores. In competitive ad hoc teamwork evaluations outside of RoboCup, more general advice would be to seek out situations that minimize the chance of poor judge scores while maximizing the chance of high judge scores.

Section 5.2.2 revealed that multiple drop-in players attempted to solely play the ball at the 2015 SPL Drop-in Player Competition. As the competition advances, we expect that (1) players who solely play the ball will perform worse and worse and (2) less players will continue to always play the ball. However, successful drop-in players must currently expect that some of their teammates may be (1) not communicating useful information, (2) not utilizing information communicated by teammates, and (3) always adopting one particular behavior, such as playing the ball. One way to handle this situation would be to use visual cues in addition to communicated information when deciding what roles teammates are fulfilling. Another approach would be to actively seek out particular roles, such as goalkeeper and defensive roles. Seeking out these roles and doing a solid job fulfilling them may be a good strategy for some drop-in players. This may be especially true when no other teammate has expressed interest in these roles and the player believes it may have weaker low-level skills than its teammates.

In a broader sense, there are plentiful research opportunities surrounding how to be a good ad hoc teammate. These research questions include where to position on the field, what types of role suggestions to communicate, and how to determine which teammates are potentially willing to coordinate.

7.3 Organizational improvements

From an organizational standpoint, organizers need to design the competition and scoring scheme such that the competition (1) is interesting to teams and (2) encourages teams to design drop-in players that are good teammates. In general, we define a good teammate as one that (1) communicates and accepts communication, (2) attempts to pass and receive passes when appropriate, (3) accepts roles and positions based on those of other teammates. We do not want drop-in players that always chase the ball to score better than those that attempt to be good teammates.

Towards the goal of encouraging the design of drop-in players that attempt to be good teammates, scoring metrics must be designed that reward good teamwork. This can be very difficult due to the issues discussed in Sects. 4 and 7.1. Adding less subjective metrics is attractive—one idea along this path for the SPL Drop-in Player Competition is an automated player tracking system that can evaluate a player's communicated information as well as its positioning. Such a system would require substantial implementation effort by league volunteers but these efforts might be decreased by using an overhead localization system similar to what is currently utilized in the RoboCup Small Size League [26].

For the 2016 SPL Drop-in Player Competition, we have multiple suggestions. First, utilize three high quality and experienced judges for all matches. Although this would require substantial time commitments from these three individuals, it is likely that the resulting judge scores would be more meaningful as a result of the improved consistency. Second, continue to utilize matches of one 10 min half, as this provides a long enough playing time to obtain meaningful results while also allowing many games to be run. Third, attempt to improve scoring metrics such that they (1) better reward players who adopt intelligent positions that are away from the ball but helpful for the team and (2) more heavily penalize robots who steal the ball from teammates or push teammates.

We also have some general suggestions for iteratively improving other competitive ad hoc teamwork evaluations. First, constantly consider how to attract and retain participants. Co-locate the competition with a multi-agent systems conference, tutorial, workshop, or summer school in order to both publicize the conference and make it easier for participants to attend. Second, analyze the scoring metric and how it affected the results in previous competitions. If the scoring metric was found to reward the wrong qualities, iteratively update it before the next competition. Finally, consider how the competition can be benchmarked over time. Benchmarking allows researchers to evaluate the progress of the competition over time.

7.4 Applications to ad hoc teamwork research

The Drop-in Player Competition was created to introduce the RoboCup community to ad hoc teamwork and take ad hoc teamwork research from individual labs to a large-scale robotics experiment. Some robotics researchers are particular interested in ad hoc teamwork because it may be the most feasible route towards creating impromptu teams of expensive, large, or otherwise difficult to obtain robots. Although the SPL Drop-in Player Competition focuses on robot soccer, being able to create functional ad hoc teams quickly is potentially helpful for many other domains—such as search and rescue or disaster recovery.

Some of the experiences shared in this article are applicable to other application domains involving ad hoc teamwork. Specifically, in the following sections we discuss (1) what ad hoc teamwork researchers can learn from our experiences and (2) how the ad hoc teamwork strategies utilized in the Drop-in Player Competition could be extended for use in other domains.

7.4.1 Ad hoc teamwork experiences

Researchers from other application domains can learn from experiences in the Drop-in Player Competition. Although some of these lessons may also be learned in smaller scale experiments, many were unexpected before running our large-scale competition.

It was expected that teammates might send noisy or incomplete information, but we quickly learned that some teammates even send consistently incorrect information. Hence, we learned that it is not only important but critical to evaluate the reasonableness of the information sent by teammates. It can be useful to track communicated information over time in order to learn a model of each teammate and calculate estimates of their trustworthiness and ability.

We initially expected that all teammates would attempt to coordinate. However, we found that when there is uncertainty about the abilities and/or trustworthiness of teammates—and/or a lack of use of communication—the natural thing to do is be self-centered and do what can be done as an individual to help the team. In fact, we often see this in humans, in the form of team projects where one or two people do most of the work or in pick-up basketball games where one player attempts to carry the team. Hence, it is important to identify which teammates are behaving in this manner. If these teammates are competent, it may be best for the overall team to support these robots in their self-centered behavior.

Since not all teammates are willing to coordinate, it becomes necessary to indicate (and notice) when an agent is willing to coordinate. In Drop-in Player Competition games, a robot might indicate willingness to coordinate through intelligent positioning, passing, or adaptive role suggestions. In other ad hoc teamwork domains, coordination may be indicated in a different manner. The important lesson to learn is that agents willing to coordinate must actively show their willingness and also notice when other agents seem willing to coordinate.

It was initially expected that some of the published ad hoc teamwork research could be directly applied when creating strategies for drop-in players. However, as we discuss further in Sect. 8, the gap between theory and practice is still large mainly due to assumptions made in most theoretical ad hoc teamwork research that do not hold up in real-life experiments with unknown robot teammates.

Finally, in domains in which heterogeneous robots are used, ad hoc teamwork can actually be easier because a robot's potential capabilities may be visually apparent based on its sensors and manipulators. Our competition did not facilitate this since our robots are physically homogeneous, but many other domains inherently have different robots with different capabilities.

7.4.2 Ad hoc teamwork strategies

Particular aspects of the strategies discussed in Sect. 5 can possibly be extended to be used in other application domains, including non-competitive domains such as search and rescue.

Work towards determining the trustworthiness of teammates could be extended to other domains in which little to nothing is known about the teammates ahead of time. Especially when self-localization or object detection is difficult, evaluating the information sent by teammates is critical.

Most ad hoc teams will need to negotiate roles either explicitly or implicitly. Some of the role negotiation metrics would be applicable to other domains. In particular, it is important to determine when a teammate is responding to the negotiation process, when a teammate is ignoring the negotiation process, and when a teammate is responding but insistent on performing a particular role.

Finally, the work on estimating the capability and reliability of teammates is clearly applicable to many other domains. Determining the capability of teammates is critical to determining both what role an agent should perform as well as what roles should be suggested for other agents. On the other hand, determining the reliability of teammates is important for determining how much to trust that the agent will accomplish the task it attempts. If a task is important, unless a teammate claiming to work on the task is both competent and reliable, it might be worth ensuring that multiple agents claim to be working on that task.

8 Related work

Ad hoc teamwork is a relatively new research area, and the Drop-in Player Competition is one of the first competitive ad hoc teamwork evaluations. In this section, we discuss the most related work that has not already been discussed in earlier sections of this article.

Although multiagent teamwork is a well-studied area, most research addresses the problem of coordinating and communicating among teams designed to work together. Indeed most RoboCup team competition entries are explicitly programmed to work together in a tightly coupled manner. STEAM is one well-know multiagent teamwork algorithm in which team members create a partial hierarchy of joint actions and monitor the current state of their plans [23]. Grosz and Kraus present a reformulation of SharedPlans in which each agent communicates its intents and beliefs and the team uses this information to coordinate joint actions [5]. Both of these architectures provide effective multiagent coordination protocols, but they require all coordinating agents to share a common coordination framework.

Some multiagent teams are designed to work specifically with their teammates in predefined ways, such as via 'locker-room agreements' [22]. Specifically, a 'locker-room agreement' is formed when there is a team synchronization period during which a team can coordinate their teamwork structure and communication protocols. The Drop-in Player Competition differs from this work in that drop-in players are not able to assume the availability of a team synchronization (pre-coordination) period. Jones et al. performed an empirical study of dynamically formed teams of heterogeneous robots in a multirobot treasure hunt domain [6]. In this work, they adapted the Traderbots system [4] to dynamically form heterogeneous teams. Their approach required a central controller and tight coordination, neither of which can be assumed in the Drop-in Player Competition.

In a 2010 AAAI challenge paper, Stone et al. challenged the artificial intelligence community to develop agents that are able to join previously unfamiliar teammates to complete cooperative activities [21]. Although they were not the first to consider this problem—one earlier work [3] is discussed below—they did draw attention to this under-researched part of multi-agent systems and they coined the terminology 'ad hoc teamwork' to describe work in this area. Their paper provided a definition of ad hoc teamwork, a methodology for evaluating performance of various ad hoc agents when paired with various teammates in a particular domain, and an initial assessment of the potential technical challenges that should be addressed when creating an ad hoc agent.

In what is, to the best of our knowledge, the first Ph.D. thesis on ad hoc teamwork, Liemhetcharat considered (1) how to model how well teammates work together on an ad hoc team, (2) how to learn such models, and (3) how to use this knowledge to form more effective ad hoc teams [9]. Liemhetcharat formally defined a weighted synergy graph that models the capabilities of robots in different roles and how different role assignments affect the overall team value. He presented a team formation algorithm that can approximate the optimal role assignment policy given a set of teammates to choose from and a task. He also used observations of a team's performance and attempted to fit models to this data, where the data could either be provided all at once (if previous observations are available) or online (to update the model as observations are acquired). This work does consider adding ad hoc agents to teams, but generally tries to answer the question of which ad hoc agents should be added to a team given multiple possible options. Drop-in players must instead determine how to behave given a set of unknown teammates.

In another ad hoc teamwork thesis, Barrett considered how to use limited knowledge about teammates to plan how to best act [2]. Barrett focused on algorithms that allowed ad hoc agents to learn about their environment and teammates, as well as reason about teamwork and choose appropriate actions. He created ad hoc agents that were (1) robust to a variety of teammates by being able to learn about teammates and adapt, (2) robust to a variety of tasks by being able to adapt to new tasks and explore teammate behaviors, and (3) able to adapt quickly to new teammates and tasks without extensive observations of either. As such, Barrett created ad hoc agents that could work well, but not necessarily optimally, with a variety of unknown teammates on a variety of tasks. Similarly to Barrett, Albrecht considered how to design an agent that is able to achieve optimal flexibility and efficiency within a team despite having no prior coordination [1]. One of his main contributions is the Harsanyi-Bellman Ad Hoc Coordination algorithm which uses concepts from game theory to facilitate ad hoc agents coordination with previously unknown agents. Both Barrett and Albrecht's methods could likely be applied to create drop-in players, although both might experience difficulties since (1) there is no learning phase, so all learning would need to occur online in real-time and (2) drop-in players initially have no information about their teammates and only limited, noisy information can be learned online during each game.

Although ad hoc teamwork is a relatively new field, there have already been some interesting theoretical results on the topic as well. For example, Wu et al. [24] proposed an online planning algorithm that works in real-time. Their algorithm selects one action at a time for each time step and performs a forward search that considers the strategies of teammates and reasons about all possible outcomes. Stone et al. [20] considered collaboration between two players. Specifically, the ad hoc agent would try to influence its teammate to attain the optimal joint utility. However, they assumed that the ad hoc agent possesses additional knowledge about the environment and knows the fixed strategies of its teammate. Since agents in the Drop-in Player Competition do not know the strategies of teammates—and it is difficult to even approximate these strategies—simulating possible outcomes or calculating optimal joint utility will be difficult if not impossible. In general, the gap between theory and practice is still large when it comes to ad hoc teamwork research.

In the robot soccer domain, Bowling and McCracken [3] propose methods for coordinating an agent that joins an unknown, pre-existing team. In their work, each ad hoc agent is given a playbook that differs from the playbook of its teammates. The teammates assign the ad hoc agent a role, and then react to it as they would any other teammate. The ad hoc agent analyzes which plays work best over hundreds of simulated games, predicts the roles its teammates will adopt in new plays, and assigns itself a complementary role in these new plays.

The RoboCup Small Size League has held a *11-vs-11 Mixed Team Challenge* [11] in which two teams are randomly combined to play as one large team consisting of 11 robots. However, the SPL Drop-in Player Competition is different from this competition in that participation in the SPL competition was much greater and at most one robot from each team joins SPL drop-in games (and hence no pre-coordination between teammates is possible). In addition, the Small Size League is based on a global vision system, enabling each robot to always observe all other participating robots.

Finally, as was mentioned earlier in this article, the RoboCup 2D simulation and 3D simulation leagues have held drop-in player technical challenges. Both leagues held initial technical challenges in 2013 along with the SPL, and the details of these challenges were published by MacAlpine et al. [10]. The 2D simulation league only held the challenge in 2013, while the 3D simulation league has continued to hold the challenge in 2014 and 2015 as well. The SPL Drop-in Player Competition is different from these technical challenges in that participation is greater, the competition is benchmarked each year by playing the team competition champion, and notably (especially given the relative ease of running simulation games) more games are run.

9 Conclusion

Despite being a relatively young competition, the SPL Drop-in Player Competition has made great strides in becoming a useful testbed for cooperation without pre-coordination. With the SPL being a standard platform league, and with options existing for teams to just compete in the SPL Drop-in Competition at RoboCup, this testbed is open and approachable for multiagent systems researchers looking to work on ad hoc teamwork in a robotics domain. The authors of this article, as well as the SPL as a whole, plan to continue this competition for the foreseeable future. The Drop-in Player Competition goal is to be able to create a team comprised of the top five drop-in players that can play comparably to the SPL main competition champion team. Such an ad hoc team may be the type of team that eventually accomplishes the RoboCup goal of beating the World Cup champion by 2050.

This article reports on the first three SPL Drop-in Player Competitions held in 2013, 2014, and 2015, suggests improvements to the competition, and provides advice for organizing new competitive ad hoc teamwork evaluations. We expect that as teams consider the strategies utilized in these three competitions and continue to work on their own entries, drop-in player strategies will continue to improve iteratively year after year. Over time, we expect research towards ad hoc teamwork in competitive ad hoc teamwork evaluations, including the SPL Drop-in Player Competition, to flourish alongside the growth of this new domain.

Acknowledgements Katie Genter and Peter Stone are part of the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (CNS-1330072, CNS-1305287), ONR (21C184-01), AFRL (FA8750-14-1-0070), and AFOSR (FA9550-14-1-0087).

References

1. Albrecht, S. V. (2015). Utilising policy types for effective ad hoc coordination in multiagent systems. Ph.D. thesis, The University of Edinburgh, Edinburgh.
2. Barrett, S. (2014). Making friends on the fly: Advances in ad hoc teamwork. Ph.D. thesis, The University of Texas at Austin, Austin, TX.

3. Bowling, M. & McCracken, P. (2005). Coordination and adaptation in impromptu teams. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI'05)*, Pittsburgh, PA.
4. Dias, B. (2004). Traderbots: A new paradigm for robust and efficient multirobot coordination in dynamic environments. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.
5. Grosz, B. J., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, *86*(2), 269–357.
6. Jones, E., Browning, B., Dias, M. B., Argall, B., Veloso, M. M., & Stentz, A. T. (2006). Dynamically formed heterogeneous robot teams performing tightly-coordinated tasks. *Proceedings of the 2006 IEEE International Conference on Robotics and Automation (ICRA'06)* (pp. 570–575), Orlando, FL.
7. Kitano, H., & Asada, M. (1998). RoboCup humanoid challenge: That's one small step for a robot, one giant leap for mankind. *Proceedings of the 1998 IEEE/RSJ International conference on intelligent robots and systems (IROS'98)* (pp. 419–424), Victoria, BC.
8. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., & Osawa, E. (1997). Robocup: The robot world cup initiative. *Proceedings of the first international conference on autonomous agents, AGENTS '97* (pp. 340–347). ACM, New York.
9. Liemhetcharat, S. (2013). Representation, planning, and learning of dynamic ad hoc robot teams. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA.
10. MacAlpine, P., Genter, K., Barrett, S. & Stone, P. (2014). The RoboCup 2013 drop-in player challenges: Experiments in ad hoc teamwork. In *Proceedings of the 2014 IEEE/RSJ international conference on intelligent robots and systems (IROS'14)*, Chicago, IL.
11. RoboCup Small Size Robot League: Small Size League/RoboCup 2015/Technical Challenges (2015). Retrieved from http://wiki.robocup.org/wiki/Small_Size_League/RoboCup_2015/Technical_Challenges.
12. RoboCup Technical Committee: Technical challenges for the RoboCup 2013 Standard Platform League competition (2013). Retrieved from http://www.tzi.de/spl/pub/Website/Downloads/Challenges2013.pdf.
13. RoboCup Technical Committee: 2014 drop-in player strategies (2014). Retrieved from http://www.tzi.de/spl/bin/view/Website/DropinStrat2014.
14. RoboCup Technical Committee: RoboCup Standard Platform League (NAO) rule book (2014). Retrieved from http://www.tzi.de/spl/pub/Website/Downloads/Rules2014.pdf.
15. RoboCup Technical Committee: 2015 drop-in player strategies (2015). Retrieved from http://www.tzi.de/spl/bin/view/Website/DropinStrat2015.
16. RoboCup Technical Committee: RoboCup Standard Platform League (NAO) rule book (2015). Retrieved from http://www.tzi.de/spl/pub/Website/Downloads/Rules2015.pdf.
17. Röfer, T., Laue, T. (2014). On B-Human's code releases in the standard platform league—software architecture and impact. In *RoboCup 2013: Robot Soccer World Cup XVII*, Lecture Notes in Artificial Intelligence, vol. 8371, pp. 648–656. Berlin: Springer.
18. Röfer, T., Laue, T., Müller, J., Schüthe, D., Böckmann, A., Jenett, D., Koralewski, S., Maaß, F., Maier, E., Siemer, C., Tsogias, A. & Vosteen, J. B. (2014). B-human team report and code release 2014. Retrieved from http://www.b-human.de/downloads/publications/2014/CodeRelease2014.pdf.
19. Röfer, T., Laue, T., Richter-Klug, J., Schünemann, M., Stiensmeier, J., Stolpmann, A., Stöwing, A. & Thielke, F. (2015). B-Human team report and code release 2015. Retrieved from http://www.b-human.de/downloads/publications/2015/CodeRelease2015.pdf.
20. Stone, P., Kaminka, G., Kraus, S., Rosenschein, J., & Agmon, N. (2013). Teaching and leading an ad hoc teammate: Collaboration without pre-coordination. *Artificial Intelligence*, *203*, 35–65.
21. Stone, P., Kaminka, G. A., Kraus, S. & Rosenschein, J. S. (2010). Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI'10)*, Atlanta, GA.
22. Stone, P., & Veloso, M. (1999). Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork. *AIJ*, *110*(2), 241–273.
23. Tambe, M. (1997). Towards flexible teamwork. *Artificial Intelligence Research*, *7*(1), 83–124.
24. Wu, F., Zilberstein, S. & Chen, X. (2011) Online planning for ad hoc autonomous agent teams. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI'11)* (pp. 439–445), Barcelona.
25. Wurman, P. R., D'Andrea, R., & Mountz, M. (2008). Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI Magazine*, *29*(1), 9–19.
26. Zickler, S., Laue, T., Birbach, O., Wongphati, M., & Veloso, M. (2010). Ssl-vision: The shared vision system for the robocup small size league. *RoboCup 2009: Robot Soccer World Cup XIII* (Vol. 5949, pp. 425–436), Lecture Notes in Computer Science Berlin Heidelberg: Springer.