Competitions for Benchmarking

Task and Functionality Scoring Complete Performance Assessment

By Francesco Amigoni, Emanuele Bastianelli, Jakob Berghofer, Andrea Bonarini, Giulio Fontana, Nico Hochgeschwender, Luca Iocchi, Gerhard K. Kraetzschmar, Pedro Lima, Matteo Matteucci, Pedro Miraldo, Daniele Nardi, and Viola Schiaffonati

Date of publication: 11 September 2015

Scientific experiments and robotic competitions share some common traits that can put the debate about developing better experimental methodologies and replicability of results in robotics research on more solid ground. In this context, the Robot Competitions Kick Innovation in Cognitive Systems and Robotics (RoCKIn) project aims to develop competitions that come close to scientific experiments, providing an objective performance evaluation of robot systems under controlled and replicable conditions. In this article, by further articulating replicability into reproducibility and repeatability and by considering some results from the 2014 first RoCKIn competition, we show that the RoCKIn approach offers tools that enable the replicability of experimental results.

Robotic Competitions and Challenges

Within the debate about the development of rigorous experimental methodologies in robotics research, the robotic competitions have emerged as a way to promote comparison of different algorithms

and systems, allowing for the replication of their results [3], [19]. Experiments and competitions present differences: an experiment evaluates a specific hypothesis, while a competition usually evaluates general abilities of robot systems. Moreover, the competitions often push the development of solutions, while experiments aim to explore phenomena and share the knowledge acquired through their results. However, merging these complementary approaches can lead to the development of an approach to competitions that makes them more scientifically grounded and suitable for benchmarking. The research and infrastructures for competitions that the robotics community has developed during the past few years can be exploited to make experimental methods in robotics sounder and more systematic, building on the common traits shared by experiments and competitions. The competitions involve robots in dynamic but controlled environments, and, having clear measures of success, these environments provide opportunities to evaluate different approaches against each other and over years of progress. Furthermore, they require integrated implementation of complete robot systems, promoting a new experimental paradigm that complements the traditional paradigm of evaluating specific modules in isolation. The competitions can thus provide a common ground for rigorously comparing different solutions, playing the role of experiments and exploiting their distinctive features, such as being appealing (both to researchers and to the general public), taking place with regularity and precise timing, promoting critical analysis of experiments out of labs, and sharing the cost and effort of setting up complex experimental installations among participants.

The robotics competitions and challenges have gained popularity from the 1970s, and now there are countless events per year.

The robotics competitions and challenges have gained popularity from the 1970s, and now there are countless events per year. From the very beginning, it has been recognized that the competitions can serve several, often conflicting, purposes, including promoting education and research to push the field forward, entertaining general audiences, and building community [5]. Although balancing these goals is sometimes difficult and some warnings have been issued about being careful not to confuse a competition with research [7], a recent trend advocates for recasting robotics challenges and competitions as experiments [3] and benchmarks [4]. Adopting this view, several competitions are currently trying to provide ways to compare the performance of different robot systems. For instance, in the field of home-assistant robots, the RoboCup@Home competition [13] evaluates robot systems in domestic environments. In the field of urban search and rescue, the Multi Autonomous Ground-Robotic International Challenge [14], the RoboCup Rescue Robot League [17], and the Defense Advanced Research Projects Agency Robotics Challenge [1] assess and measure the capabilities of different types of robots in real disaster environments.

The approach promoted by the RoCKIn project (http://rockinrobotchallenge.eu) aims to move from competitions that provide benchmarking at the system-level based on a single high-level measure to more sophisticated benchmarking activities. The RoCKIn competitions come close to scientific experiments as they provide a rigorous performance evaluation of robot systems under controlled and reproducible conditions. More precisely, the competitions adopt the classical conceptual framework of scientific experimental methods that separates reproducibility from repeatability [11] and using the results from the first RoCKIn competition, we show how RoCKIn can provide a set of tools to enable the replicability of experiments involving autonomous robots.

Replicability: Reproducibility and Repeatability

Among the principles that characterize the scientific experimental method, replicability is considered fundamental to allow for rigorous comparison of results and thus affects the processes and products of scientific research. The concept of replicability has emerged as central to the debate within autonomous robotics to make its methods closer to the standard of rigor of other scientific disciplines [6]. Usually, the replicability in robotics research is intended as the possibility to reproduce published results. However, the issue is more complex and problematic, as it has been recognized in other fields of computer science [9]. Therefore, to better articulate this concept and the contributions of the RoCKIn approach, we take into account the traditional conceptualization as devised in the history and the philosophy of science. Accordingly, the replicability can be specified into reproducibility and repeatability. Although they both refer to the general idea that scientific results should undergo the most severe criticisms to be strongly confirmed, they indeed point out two distinct characteristics of experimental methodology [11].

- *Reproducibility* is the possibility to verify, in an independent way, the results of a given experiment. It refers to the fact that other experimenters, different from the ones claiming validity for some results, are able to achieve the same results by starting from the same initial conditions, using the same type of instruments and parameters, and adopting the same experimental techniques. To be reproducible, an experiment must be fully documented.
- *Repeatability* concerns the fact that a single result is not sufficient to ensure the success of an experiment. A successful experiment must be the outcome of a number of trials, possibly performed at different times and in different places. These requirements guarantee that results have not been achieved by chance but are systematic, and that statistically significant trends can be identified.

How can these two very general features be applied in the practice of robotics research, and, in particular, of robotics competitions?

To answer this question, we force an artificial separation between the two intertwined concepts. Concerning reproducibility, the need for a precise description of results and of the processes adopted to achieve those results calls for the following requirements:

- conducting reproduced experiments in the same settings of the original ones that, therefore, should be explicitly and fully specified to be exactly reproduced
- making the used code and data available to the research community; note, however, that the mere availability of code and data does not guarantee reproducibility. The source code and test data have to be available, the code has to build, the execution environment has to be replicated (including the robot system or part of it), the code has to run to completion, and accurate measurements have to be collected [8].

Concerning repeatability, the mere repetition of runs is surely a way to attain repeatability, but it is only the first step in the direction of achieving systematic results. Given that one of the goals in making experiments is to obtain generalizations, repeatability can be practically achieved by the following:

- conducting a serious analysis of how many runs of a robotic experiment are required to obtain statistically significant results [16]
- performing experimental sessions that take place in settings S' that are fully compatible with the description of original settings S, but that might slightly differ for some details left unspecified in S. This contributes to filter out casual issues that affect experimental outcomes.

The enactment of the above experimental requirements to competitions is the basis of the approach followed in RoCKIn.

Overview of the RoCKIn Approach

The RoCKIn project aims to provide tools for benchmark-ing to the robotics community by designing and setting up competitions that increase scientific and technological knowledge. The RoCKIn competitions retain the tradition-al value of producing a ranking among alternative solutions at competition time, assigning prizes and awards to the best teams, and stimulating progress. At the same time, the ex-perimental settings of the competitions gain a more general significance as benchmarking procedures. The RoCKIn project moves from competitions providing benchmarking with a single system-level measure (like the score of a soc-cer game) to a more sophisticated benchmarking approach integrated within competitions, where different elements are evaluated and benchmarking results can be used not only to rigorously compare robot systems, but also to better understand them. According to this perspective, we could say that the RoCKIn competitions come close to scientific experiments by providing an objective performance evalua-tion of a robot system/subsystem under controlled and re-producible conditions.

Two challenges have been selected as competition scenari-os in this project due to their high relevance and impact -on Europe's societal and industrial needs: domestic service robots (RoCKIn@Home) and innovative robot applications in industry (RoCKIn@Work). Both challenges have been inspired by similar activities in the RoboCup community [15], [20]. The RoCKIn aims at exploiting some of the RoboCup achievements to extend the pure competition approach in several aspects, as summarized in Table 1.

In RoCKIn@Home [18], Granny Annie lives in an apartment together with some pets and presents some of the typical problems of aging people. The aim of RoCKIn@Home is to develop robots that support Granny Annie and her quality of life. The RoCKIn@Home test bed reflects an ordinary European apartment with all common household items like windows, doors, furniture, and decorations.

The RoCKIn@Work scenario [10] represents a medium-sized factory that specializes in the production of small- to medium-sized lots of mechanical parts and assembled mechatronic products, which tries to optimize its production process to meet the increasing demands of their customers. This factory thus requires a system with two essential capabilities: 1) mobile manipulation to perform tasks such as assembly processes, quality controls, order handling, and logistics and 2) autonomy in switching between different tasks. The RoCKIn@Work test bed also includes networked devices such as force fitting machines and conveyor belts, which can be operated by the robots themselves.

The RoCKIn competitions come close to scientific experiments as they provide a rigorous performance evaluation of robot systems under controlled and reproducible conditions.

One of the main features of the RoCKIn competitions is the introduction of two separate classes of evaluations, task benchmarks (TBMs) and functionality benchmarks (FBMs). The former are devoted to evaluating the performance of integrated robot systems, while the latter focus on the performance of specific subsystems (like object recognition and

Table 1. The shift from RoboCup to RoCKIn.

From RoboCup …	… To RoCKIn
Adopts a pure competition approach with a (mostly) monofaceted scoring of tasks	Adopts a more sophisticated competition approach with multifaceted scoring of both tasks and functionalities
Does not explicitly address benchmarking	Explicitly considers structured and repeatable benchmarking
Presents mostly passive environments	Integrates sensors and actuators in the environment and wirelessly networks them with mobile robots

speech understanding). A TBM deals with complete robot systems, implying that a large set of interacting robot subsystems (like navigation, perception, and manipulation) are examined together at the same time. FBMs, on the other hand, focus on the performance of single subsystems, defining a precise setup in which a single robot functionality can be evaluated. This evaluation is performed according to well-specified quantitative measures and criteria, which depend on the functionality being tested. The scoring of TBMs and FBMs is then used to determine rankings of teams and to award prizes at the competitions.

The RoCKIn approach builds on the efforts of the RoboCup community to identify which functionalities are stable and solved or unsolved, at least in the context of the selected competitions. However, the evaluation of these functionalities in the RoboCup is mixed with the evaluation of the tasks, and separating the two is difficult. "In fact, teams obtained good results in navigation, mapping, person tracking, and speech recognition (with the average above 50%, except for navigation). Notice that the reason for a low-percentage score in navigation is not related to inabilities of the teams, but it is because part of the navigation score is only available after some other task was achieved [13]." The RoCKIn approach avoids this problem by limiting the influence of all other subsystems when evaluating a robot functionality in a FBM. For example, for testing object perception, robots are put in place before starting the test. More generally, we mitigate the difficulty of separating subsystems under investigation from their environments, which also include other robot subsystems that are not being evaluated, by carefully designing FBMs that minimally involve these last subsystems.

The replicability can be specified into reproducibility and repeatability.

The TBMs and FBMs are evaluated differently. The TBMs measure the achievement of goals, which is a yes or no answer to specific questions (e.g., "Does the robot understand Annie's command(s)? Does it correctly identify the requested object?"). The FBMs measure robot performance, which is a number resulting from the measures used for scoring and ranking, such as effectiveness (e.g., precision and recall) and efficiency (time, resources used, and so on), as further discussed in the next section. This division resembles the recently proposed evaluation of artificial intelligence systems [12], which is based on task-oriented and ability-oriented evaluations. In both the cases, the RoCKIn approach tries to avoid subjective evaluation to improve reproducibility. Indeed, attention has been dedicated to requirements for benchmarking and scoring runs as autonomously as possible (i.e., without continuous human intervention), specifically by using automated computing systems called *RefBoxes* (also called *Central Factory Hub* in RoCKIn@Work). In this respect, the more automated scoring approach of RoCKIn contrasts, for instance, with that of the RoboCup Rescue Robot League [17], which is heavily based on human judges.

In addition to ranking teams in the competitions, the approach of RoCKIn promises to be a good way to understand robot systems because it enables researchers to study the impact of functionality performance on task performance. Moreover, it forces teams to develop means of continuously monitoring the performance of their robot systems because they have to provide regular feedback to the RefBoxes and store data for benchmarking.

Table 2. The teams participating in the 2014 RoCKIn competition.

RoCKIn@Home Teams
b-it-bots@Home, Bonn-Rhein-Sieg University of Applied Sciences, Germany
BARC, University of Birmingham, United Kingdom
Homer@UniKoblenz, University of Koblenz-Landau, Germany
Pumas@Home, Universidad Nacional Autonoma de Mexico, Mexico
SocRob@Home, Universidade de Lisboa, Portugal
UrsusTeam, University of Extremadura, Spain
Watermelon Project, University of Leon, Spain
RoCKIn@Work Teams
b-it-bots@Work, Bonn-Rhein-Sieg University of Applied Sciences, Germany
IASLab@Work, University of Padua, Italy
SPQR@Work, Sapienza University of Rome, Italy

The First RoCKIn Competition

The first RoCKIn competition (http://rockinrobotchallenge.eu/rockin2014.php) was held in Toulouse, France, 26–30 November 2014, and was the first opportunity to test the practical application of the approach outlined in the previous section and to prepare for the final RoCKIn competition (http://rockinrobotchallenge.eu/rockin2015.php) to be held at the end of 2015 in Lisbon, Portugal. The teams participating in the 2014 event are listed in Table 2.

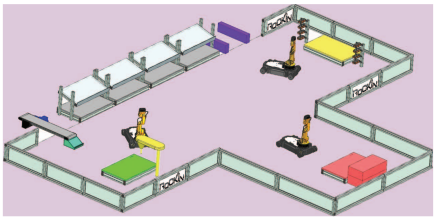
Setup for the Competition

The detailed specifications of the RoCKIn@Home and RoCKIn@Work test beds are reported in the corresponding rule books, which are available at http://rockinrobotchallenge.eu/publications.php under "deliverables and reports" and allow for their precise reproduction at other sites than those of the competitions. The layouts and the sizes of the RoCKIn@Home and RoCKIn@Work arenas (Figures 1 and 2) are fully specified, together with the materials of walls and the precise definition of objects present in the environments (at a level of detail that include furniture and floristic objects for RoCKIn@Home). The robots must conform to certain size, weight, and safety restrictions and can be wirelessly networked with other devices. Apart from this, teams are free to

(a)

(b)

Figure 1. The RoCKIn@Home arena: Granny Annie's apartment. (a) The 3-D model of the RoCKIn@Home arena and (b) the real RoCKIn@Home arena built for the first RoCKIn competition.



(a)

(b)

Figure 2. The RoCKIn@Work arena: a model of a medium-size factory.

choose the robot platforms they deem most adequate to obtain the best performance. Shortly, with the idea to attain reproducibility and repeatability, the RoCKIn precisely specifies several aspects of the settings S in which the competitions take place, but some aspects, like the sensor equipment of the robots, are left unspecified. In this way, if a robot capability is demonstrated in S and in other settings S' that differ from S for the actual implementation of the aspects unspecified in S, it can be concluded that the capability is stable or solved.

TBMs and FBMs defined for the 2014 edition of the RoCKIn@Home and RoCKIn@Work challenges are listed in Table 3.

Table 3. TBMs and FBMs for the 2014 RoCKIn competition.

RoCKIn@Home	
TBM1@Home	Getting to know my home
TBM2@Home	Welcoming visitors
TBM3@Home	Catering to Granny Annie's comfort
FBM1@Home	Object perception
FBM2@Home	Object manipulation
FBM3@Home	Speech understanding
RoCKIn@Work	
TBM1@Work	Assemble aid tray for force fitting
TBM2@Work	Plate drilling
TBM3@Work	Prepare box for manual assembly step
FBM1@Work	Object perception
FBM2@Work	Visual servoing

Scoring of TBMs

The scoring of TBMs is based on achievements and penalties. Specifically, performance classes C_n are defined for ranking robot performance in a task, based on the number (n) of achievements (A) that the robot reaches during the execution of the task. Within each performance class C_n (i.e., the number of achievements being equal), ranking is defined according to the number of penalty behaviors (PB) of the robot that represent errors while executing the task. More formally, given a task, the following rules are applied:

- the ranking of any robot belonging to performance class C_n is considered better than that of any robot belonging to performance class C_m with $m < n$; class C_0 is the worst performance class
- among robots belonging to the same performance class, the robot which received fewer penalties is considered higher in rank
- among robots belonging to the same performance class and with the same number of penalties, the ranking of the robot that completed the task in a shorter time is considered higher.

Moreover, to ensure the safety of the competition, disqualifying behaviors (DB) are defined, namely the things that a robot must not do to avoid being excluded from the competition. For security reasons, a human referee always has access

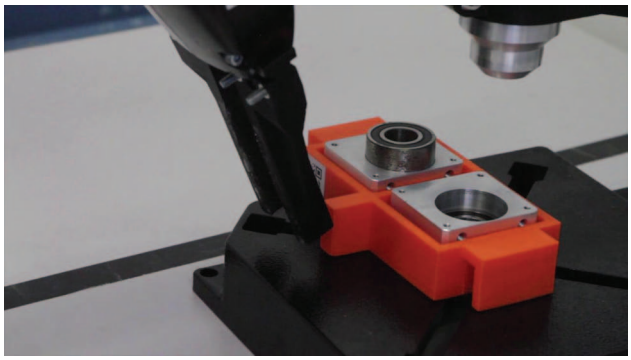


Figure 3. A robot assembling an aid tray for force fitting (TBM1@Work). The picture comes from the RoCKIn consortium.

Figure 4. The objects used in the FBM1@Home object perception. The picture comes from the RoCKIn consortium.

Table 4. The results of the FBM1@Home object perception.

Team	Object Class Accuracy
UrsusTeam	0.90
Homer@UniKoblenz	0.80
Pumas@Home	0.70
Watermelon Project	0.30

to a red button that, if pushed, stops the robot in case of disqualifying behaviors.

Each team had the possibility of performing five runs of each TBM in RoCKIn@Home (two runs on the first day, two runs on the second day, and one run on the last day) and three runs of each TBM in RoCKIn@Work (one run per day). The best score over all runs determined the winner for the TBM.

The recorded data depend on the hardware equipment of the robots.

One key property of this scoring system is that a robot that executes the task completely will always be ranked better than a robot that executes the task partially. Penalty behaviors do not change the performance class of a robot and only influence intraclass ranking. It is also possible to envisage the use of weighted penalties; however, this makes the ranking criteria harder to understand and apply. Therefore, weights have not been used in the 2014 RoCKIn competition.

Sets A, PB, and DB are task-dependent. For example, for the TBM3@Home, catering to Granny Annie's comfort, the sets are as follows (other TBMs sets are defined similarly, please refer to the rule books for full details).

- A = {upon reception of a call signal, the robot enters the room where Granny Annie is waiting, the robot understands the commands uttered by a person playing the role of Granny Annie and the robot operates the right devices requested by Granny Annie, the robot finds the right objects, the robot brings Granny Annie the right objects}.
- PB = {the robot bumps into furniture, the robot drops an object, the robot stops working}.
- DB includes, for example, hitting Granny Annie.

Scoring TBMs in the RoCKIn@Work is performed similarly. For instance, achievements for the TBM1@Work assemble aid tray for force fitting task include the correct identification of the assembly aid tray and the correct delivery of the aid tray to the force fitting machine (see Figure 3). Similarly, PBs are defined, like dropping an object or bumping into obstacles.

Overall, the performance of the teams in the TBMs of the first RoCKIn competition has been good for the achievements related to navigation, while other achievements have proved to be more challenging. Full results are available at http://rockinrobotchallenge.eu/rockin2014.php under "Results."

Scoring of FBMs

The scoring of FBMs measures the performance of robot subsystems and is specific for each functionality. For example, consider the FBM1@Home object perception (other FBMs are scored similarly), which is used to assess the capabilities of a robot in identifying the class (e.g., cups), the instance (e.g., black coffee mug), and the pose (with respect to a global coordinate system) of objects that are presented to it and that are relevant to the TBMs of RoCKIn@Home (Figure 4 shows a sample of the possible objects). Each team that participated in this FBM had the opportunity to perform the benchmark four times, each time trying to identify ten randomly selected objects placed in random poses on a flat table (which was located at a fixed and known pose in the global coordinate system). The scoring considers the accuracy in class classification (and, in case of tie, the accuracy in instance classification, the error rate in identifying the three-dimensional pose of the object, and the test time, in this order). The best score over the four runs is considered for the final ranking. Final results for the FBM1@Home object perception are reported in Table 4.

The FBM3@Home speech understanding has the goal of evaluating the ability of robot systems to understand speech commands that a user (like Granny Annie) gives in a home

Figure 5. The robots during the FBM3@Home speech understanding.

environment (like go to the living room, put the jar on the table). Five teams participated in this FBM. Each team had the opportunity to perform four runs, each time trying to understand a number of command sentences provided to the robots both as audio files (between 30 and 50) on a USB stick and as sentences (from 6 to 12) spoken by a person through a microphone (to retain reproducibility, the person speaking was always the same for all teams and runs). A single loudspeaker placed on the ground and facing up was used to produce an omnidirectional audio source so that the robots were able to perform the test in parallel (see Figure 5).

The scoring considers the ability to recognize the main actions (like go and put) and the main arguments (like living room, jar, and table) of the commands. This is measured in terms of the accuracy in correctly classifying the arguments (AgC), the accuracy in correctly classifying the actions (AcC), and the word error rate (WER) in correctly recognizing each word. Ranking is based on AcC, AgC, and WER, in this order. A total of 15 valid team runs (i.e., with nonzero performance) were performed. Full results for the FBM3@Home speech understanding are reported in Table 5. The difficulty of the sentences (evaluated by an expert in speech recognition) was increasing during the first three runs, while the fourth run contained mixed sentences. This is reflected in the performance of the robots over the runs.

In a similar manner, for RoCKIn@Work, FBM1@Work (object perception) and FBM2@Work (visual servoing) are defined. The former focuses on the detection, recognition, and localization of industrial objects, where the latter focuses on controlling the manipulator motion based on its own visual perception.

Note that the scoring of TBMs and FBMs relies on the RefBoxes that support detecting achievements and penalties and measure the performance of the robots (e.g., the time spent in performing an activity). In particular, the RefBoxes can largely automate the evaluation of FBMs. For example, in the FBM1@Home object perception, the RefBox randomly selects which objects will be presented to the robots, sends the start signal to the robots, and waits for

Table 5. The results of the FBM3@Home speech understanding.

Run 1	AgC	AcC	WER
UrsusTeam	0.28	0.76	0.47
b-it-bots@Home	0	0	0.70
Homer@UniKoblenz	0	0	0.74
Run 2	**AgC**	**AcC**	**WER**
UrsusTeam	0.24	0.65	0.53
Pumas@Home	0.07	0.46	0.59
b-it-bots@Home	0.05	0.11	0.94
Homer@UniKoblenz	0	0.37	0.70
Run 3	**AgC**	**AcC**	**WER**
UrsusTeam	0.03	0.62	0.50
Pumas@Home	0.03	0.18	0.59
b-it-bots@Home	0	0.43	0.75
Homer@UniKoblenz	0	0.34	0.76
Run 4	**AgC**	**AcC**	**WER**
UrsusTeam	0.10	0.71	0.47
Pumas@Home	0.08	0.35	0.74
b-it-bots@Home	0.01	0.30	0.72
Watermelon Project	0	0	0.69

replies from the robots. Moreover, for TBMs the RefBoxes manage the communication between the test bed and the robots, mediating between the environment devices and the robots. This helps to identify if devices (like force fitting machine in RoCKIn@Work) are correctly actuated.

Benchmarking and Replicability

During the RoCKIn competitions, we plan to collect data for benchmarking that go beyond those strictly needed for scoring the runs of the robots. Benchmarking data are acquired both by the robots and by devices in the environment.

For example, in TBM3@Home, catering to Granny Annie's comfort, the following data were expected to be collected for each run of every team: the audio signals of the conversations between Granny Annie and the robot (collected by the robot), the final commands produced after the natural language analysis process (collected by the robot), the ground truth pose of the robot while moving in the environment (collected using the OptiTrack motion capture system by NaturalPoint), the pose of the robot while moving in the environment (as perceived by the robot), the sensorial data of the robot when recognizing the object to be operated, and the results of the robot's attempts to execute Granny Annie's commands.

The scoring of TBMs is based on achievements and penalties.

For the FBM1@Home object perception, expected benchmarking data include sensor data (images, point clouds, and so on) used by the robot to perform classification; the class; the

instance, and the pose of every object (as determined by the robot); and the actual class, instance, and pose of every object (ground truth). For the FBM3@Home speech understanding, benchmarking data that were expected to be collected include sensor data (audio files) used by the robot to perform speech recognition and the command (action and arguments) as recognized by the robot. Similar rich benchmarking data were expected to be collected for all other FBMs and TBMs. Note that benchmarking data include ground truth, for example, the poses of the robots and objects and the commands issued to the robots.

For the participating teams, the recording of sensor data and processed information is mandatory, although some flexibility has been allowed during the first RoCKIn competition. Since the process is rather invasive and it turns out that most teams use ROS (http://www.ros.org), we tried to limit the effort for onboard data collection by using the ROS built-in recording tool called rosbag (which can also be used by teams not using ROS, by exploiting the rosbag Application Programming Interfaces). Note that the recorded data depend on the hardware equipment of the robots. For example, data collected during the FBM1@Home object perception include both images and images plus point clouds of the same objects, according to the different sensors mounted on different robots. In principle, stereo images could also be present. The amount of benchmarking data collected over all the runs of the TBMs and FBMs on the three days of the 2014 competition is summarized in Table 6. A positive trend is evident as the competition progressed, from 43% of runs (10 out of 23 runs) with complete benchmarking data on the first day, to 91% of runs (20 out of 22 runs) with complete benchmarking data on the last day, which was a half-day competition. This is due to increased awareness about data collection. Globally, 68% of runs (52 out of 76 runs) have complete benchmarking data. Incomplete benchmarking data are due to their incorrect format or to missing portions. Note that the runs with no benchmarking data also include runs in which the robots failed to start, which were 4, 3, and 0, on the three days, respectively.

The robots must conform to a certain size, weight, and safety restrictions and can be wirelessly networked with other devices.

Table 6. The benchmarking data collected during the 2014 RoCKIn competition.

	Day 1	Day 2	Day 3
Total runs	23	31	22
Runs with complete data	10	22	20
Runs with incomplete data	2	1	1
Runs with no data	11	8	1

These benchmarking data are made available to the research community, to ease the reproducibility of results and the comparison with the teams participating in the RoCKIn competitions. The benchmarking data can be found at http://thewiki.rockinrobotchallenge.eu/. In particular, data relative to poses of robots collected by the ground truth system can be used by the teams to replay the runs of their robots, for example, matching the actual pose of a robot with the expected one according to the robot perception. As some anecdotal evidence from the 2014 competition confirmed, this can have a positive impact on fixing bugs and improving the performance of teams. The RoCKIn competitions aid in collecting a huge amount of data that can be later used for reproducing experiments and for benchmarking by researchers not participating in the competitions, as researchers can download the data sets and run their algorithms on them. For example, laser range scanner data collected during task benchmarks can also be employed to test and evaluate mapping and localization algorithms, while audio files collected during the FBM3@Home speech understanding can be used to test algorithms for speech understanding. Researchers can also compare their results with those obtained by teams in the RoCKIn competitions. In this sense, the availability of data recorded by the robots with different configurations while performing the same task or functionality benchmark enrich the data sets provided by the RoCKIn.

Some steps toward the repeatability of experiments in the context of a competition were taken: each team was given the option of repeating all the TBMs and FBMs at least three times (although some teams performed less runs due to robot failures or the decision to skip them). Selecting the best-scored run makes sense for the competition ranking (see Table 4), but the results over all the runs could be considered for a statistical analysis of the significance of the observed differences in performance. However, the data from the first RoCKIn competition are not enough to support such a statistical analysis yet. We are working on teaching the teams to use RoCKIn benchmarking infrastructure more systematically in the 2015 competition.

Conclusions and Future Works

With this article, we have pointed out how the RoCKIn approach to competitions makes them closer to replicable scientific experiments, as the benchmarking procedures we defined can provide a rigorous and articulated performance evaluation of the robot systems under controlled circumstances. By taking inspiration from the history and philosophy of science, we have articulated replicability into reproducibility and repeatability, and suggested how to apply them in the practice of robotic research. From the analysis of some results from the 2014 RoCKIn competition, we can say that the RoCKIn approach contributes to enabling the reproducibility of experimental results by providing full details to reproduce test beds and by collecting rich benchmarking

data. As for repeatability, while the structure of the RoCKIn competitions pushes in this direction, the results of the 2014 RoCKIn competition are still too preliminary to draw any conclusion.

Future work will address the situations that the current version of the RefBoxes cannot manage, like detecting if a robot has hit something or someone or has correctly grasped an object, to make scoring even more automatic. More generally, we will promote the further development of the RoCKIn approach, whose final competition is planned for the end of 2015, toward fully reproducible experiments. It is expected that enough teams will participate to get a significant amount of data that will enable a systematic and quantitative analysis of robot performance, also relative to the evaluation of the importance of single functionalities in the execution of complex tasks. In this respect, we plan to investigate the use of some tools from game theory, like Shapley values and power indexes [2].

References

[1] (2015). DARPA robotics challenge. [Online]. Available: http://www.theroboticschallenge.org

[2] A. Ahmad, I. Awaad, F. Amigoni, J. Berghofer, R. Bischoff, A. Bonarini, R. Dwiputra, G. Fontana, F. Hegger, N. Hochgeschwender, L. Iocchi, G. Kraetzschmar, P. Lima, M. Matteucci, D. Nardi, V. Schiaffonati, and S. Schneider. (2014). RoCKIn deliverable D1.2 General evaluation criteria, modules and metrics for benchmarking through competitions. [Online]. Available: http: //rockinrobotchallenge.eu/publications.php

[3] M. Anderson, O. Jenkins, and S. Osentoski, "Recasting robotics challenges as experiments," *IEEE Robot. Automat. Mag.*, vol. 18, no. 2, pp. 10–11, 2011.

[4] S. Behnke, "Robot competitions—Ideal benchmarks for robotics research," in *Proc. IROS Workshop Benchmarks Robotics Research*, 2006.

[5] P. Bonasso and T. Dean, "A retrospective of the AAAI robot competitions," *AI Mag.*, vol. 18, no. 1, pp. 11–23, 1997.

[6] F. Bonsignorio, J. Hallam, and A. del Pobil. (2007). Special interest group on good experimental methodology—GEM guidelines. [Online]. Available: http://www.heronrobots.com/EuronGEMSig/downloads/GemSigGuidelinesBeta.pdf

[7] T. Bräunl. "Research relevance of mobile robot competitions," *IEEE Robot. Automat. Mag.*, vol. 6, no. 4, pp. 32–37, 1999.

[8] C. Collberg, T. Proebsting, G. Moraila, A. Shankaran, Z. Shi, and A. Warren. (2014). Measuring reproducibility in computer systems research. [Online]. Available: http://reproducibility.cs.arizona.edu/tr.pdf

[9] C. Drummond, "Replicability is not reproducibility: Nor is it good science," in *Proc. Evaluation Methods Machine Learning Workshop 26th ICML*, 2009.

[10] R. Dwiputra, J. Berghofer, A. Ahmad, I. Awaad, F. Amigoni, R. Bischoff, A. Bonarini, G. Fontana, F. Hegger, N. Hochgeschwender, L. Iocchi, G. Kraetzschmar, P. Lima, M. Matteucci, D. Nardi, V. Schiaffonati, and S. Schneider, "The RoCKIn@Work challenge," in *Proc. ISR/Robotik*, 2014, pp. 328–333.

[11] I. Hacking. *Representing and Intervening.* Cambridge, U.K.: Cambridge Univ. Press, 1983.

[12] J. Hernández-Orallo. (2014). AI evaluation: Past, present and future. [Online]. Available: http://arxiv.org/abs/1408.6908

[13] D. Holz, L. Iocchi, and T. van der Zant, "Benchmarking intelligent service robots through scientific competitions: The RoboCup@Home approach," in *Proc. AAAI Spring Symp. Designing Intelligent Robots: Reintegrating AI II*, 2013, pp. 27–32.

[14] A. Hsieh and S. Lacroix, Eds., "Special issue on multiautonomous ground-robotic international challenge (MAGIC)," *J. Field Robot.*, vol. 29, no. 5, pp. 687–841, 2012.

[15] G. Kraetzchmar, N. Hochgeschwender, W. Nowak, F. Hegger, S. Schneider, R. Dwiputra, J. Berghofer, and R. Bischoff, "RoboCup@Work: Competing for the factory of the future," in *Proc. RoboCup Symp.*, 2014.

[16] J. Parker, J. Godoy, W. Groves, and M. Gini, "Issues with methods for scoring competitors in RoboCup rescue," in *Proc. AAMAS Workshop Autonomous Robots Multirobot Systems*, 2014.

[17] J. Pellenz, A. Jacoff, T. Kimura, E. Mihankhah, R. Sheh, and J. Suthakorn, "RoboCup rescue robot league," in *Proc. RoboCup Symp.*, 2014.

[18] S. Schneider, F. Hegger, A. Ahmad, I. Awaad, F. Amigoni, J. Berghofer, R. Bischoff, A. Bonarini, R. Dwiputra, G. Fontana, L. Iocchi, G. Kraetzschmar, P. Lima, M. Matteucci, D. Nardi, and V. Schiaffonati, "The RoCKIn@Home challenge," in *Proc. ISR/Robotik*, 2014, pp. 321–327.

[19] B. Smart, "Competitions, challenges, or journal papers," *IEEE Robot. Automat. Mag.*, vol. 19, no. 1, p. 14, 2012.

[20] T. Wisspeintner, T. van der Zant, L. Iocchi, and S. Schiffer, "RoboCup@Home: Scientific competition and benchmarking for domestic service robots," *Interaction Studies*, vol. 10, no. 3, pp. 392–426, 2009.

Francesco Amigoni, Politecnico di Milano, Italy. E-mail: francesco.amigoni@polimi.it.

Emanuele Bastianelli, University of Rome La Sapienza, Italy. E-mail: emanuele.bastianelli@gmail.com.

Jakob Berghofer, KUKA Laboratories GmbH, Augsburg, Germany. E-mail: Jakob.Berghofer@kuka.com.

Andrea Bonarini, Politecnico di Milano, Italy. E-mail: andrea.bonarini@polimi.it.

Giulio Fontana, Politecnico di Milano, Italy. E-mail: giulio.fontana@polimi.it.

Nico Hochgeschwender, Bonn-Rhein-Sieg University, Germany. E-mail: nico.hochgeschwender@h-brs.de.

Luca Iocchi, University of Rome La Sapienza, Italy. E-mail: iocchi@dis.uniroma1.it.

Gerhard K. Kraetzschmar, Bonn-Rhein-Sieg University, Germany. E-mail: gerhard.kraetzschmar@brsu.de.

Pedro Lima, Instituto Superior Técnico, University of Lisbon, Portugal. E-mail: pal@isr.tecnico.ulisboa.pt.

Matteo Matteucci, Politecnico di Milano, Italy. E-mail: matteo.matteucci@polimi.it.

Pedro Miraldo, Instituto Superior Técnico, University of Lisbon, Portugal. E-mail: pmiraldo@isr.tecnico.ulisboa.pt.

Daniele Nardi, University of Rome La Sapienza, Italy. E-mail: nardi@dis.uniroma1.it.

Viola Schiaffonati, Politecnico di Milano, Italy. E-mail: viola.schiaffonati@polimi.it.