Towards Using Multiple Iterated, Reproduced, and Replicated Experiments with Robots (MIRRER) for Evaluation and Benchmarking[⋆]

Adam Norton[1][0000−0002−6127−4588] and Brian Flynn[1][0000−0001−5549−4528]

New England Robotics Validation and Experimentation (NERVE) Center,
University of Massachusetts Lowell, Lowell, MA, 01852, USA
`adam_norton,brian_flynn@uml.edu`
`https://nerve.uml.edu/`

Abstract. The robotics research field lacks formalized definitions and frameworks for evaluating advanced capabilities including generalizability (the ability for robots to perform tasks under varied contexts) and reproducibility (the performance of a reproduced robot capability in different labs under the same experimental conditions). This paper presents an initial conceptual framework, MIRRER, that unites the concepts of performance evaluation, benchmarking, and reproduced/replicated experimentation in order to facilitate comparable robotics research. Several open issues with the application of the framework are also presented.

Keywords: Robotics · Benchmarking · Performance evaluation · Generalizability · Reproducibility · Replicability.

1 Introduction and Background

The benefit of advanced robotics over traditional automation is the ability for robotic capabilities to generalize across domains and applications; in other words, a robotic component or system has the ability to perform a task with some level of variation. This capability is referred to by many names including robustness, flexibility, versatility, and generalizability, all of which refer to the robot's ability to operate when variations are induced. Open issues with developing generalizable robotics solutions include a lack of accepted frameworks and terminology to properly define their capabilities as well as a lack of effective benchmarking methodologies. The use of learning-based algorithms in robotics allows for generalizable solutions to be developed; however, issues with reproducibility and replicability – of the experimental conditions and, more importantly, of the results – are incurred. The non-deterministic nature of these approaches coupled with the complexity of a physical robot system operating in a real-world environment (although similar issues exist in simulation) is problematic.

While issues with reproducibility and replicability are not unique to robotics research, there is a multitude of evidence in the field that they are prominent (e.g., issues with reusing code [11,12], reproducing deep reinforcement learning [23], multiple human-robot interaction specific concerns [15,20]) with some efforts towards solving them (e.g., IEEE R-articles [8,7], ACM badging to denote reproduced or replicated results [2]). Particularly in the domain of robotic manipulation, benchmarking tools have also been developed to bolster the repeatability of physical robot testing (e.g., accurately repositioning artifacts via automated mechanisms [10,14] or augmented reality [19]) and the reproducibility of research results (e.g., task protocols to benchmark performance [21,6], datasets for comparing results [25]). A survey conducted of the robotics research community and a series of workshops on the current state of open-source assets and benchmarking resources reveals that a lack of consensus on metrics, protocols, software component structures, and incentives are among the open issues impacting reproducibility and replicability [27].

If a robotic capability is purported to be generalizable, one should be able to evaluate this assertion by iterating on the conditions of an experiment and determining to what level the two varied conditions impact performance. Similarly, the reproducibility of a robotic capability should also be able to be evaluated by reproducing the conditions of an experiment (or set of experiments), which can then also be used to evaluate generalizability. Lastly, the same techniques used to reproduce an experiment should also allow for different but comparable robotic capabilities to be evaluated and compared across labs (referred to as replicability, defined later). We propose to unite these concepts under a single framework in order to formalize their definitions, articulate their interrelated nature when it comes to conducting evaluations, and justify the development of new methodologies and software/hardware architectures to improve robot benchmarking. While all issues are not addressed in this paper, an initial conceptual framework (MIRRER) is presented that unites these concepts towards solving the challenges associated with evaluating, reproducing, and replicating robotics research.

2   MIRRER Framework

Throughout this section, all examples provided use a robotic grasping experiment for a pick-and-place task as a notional scenario. We first provide proposed definitions for the concepts and terminology used in the Multiple Iterated, Reproduced, and Replicated Experiments with Robots (MIRRER) framework:

– **Context** of a robotic experiment refers to the parameters of the system's operation that can impact performance, distilled into four categories (adapted from [26]):
  - **Input data** provided to the robot to perform its task (e.g., 3D models of objects with pre-planned grasps, point clouds collected in-situ).
  - **Target objects** being interacted with (e.g., objects to be grasped, kit form obstructions to avoid colliding with).

- **Tasks** being performed with or around those objects (e.g., grasp object, pick up, and place in kit).
  - **Environment** where the task is being performed (e.g., ambient lighting, layout of bin with objects and kits).
- **Robot system** executing the tasks whose configuration consists of multiple components including perception modules, motion planners, grasp planners, hardware, etc. The component that is being evaluated through experimentation is referred as the **component under evaluation (CUE)**.
- **Generalizability** of the CUE's performance is a metric that is evaluated when multiple experiments are conducted in the same or different labs under intentionally varied contexts (e.g., light vs. dark lighting conditions).
- **Iterated** experiments are conducted in the same lab and use either:
  - Different context parameters and the same robot system configuration to evaluate **generalizability** of the CUE's performance (e.g., evaluating a grasp planner using two sets of target objects), or
  - The same context parameters and a different robot system configuration to **compare** the performance of multiple CUEs (e.g., comparing the performance of two grasp planners).
- **Reproduced** experiments are conducted in a different lab, using the same context parameters and the same robot system configuration to evaluate **reproducibility** of the CUE's performance (e.g., lab 2 reproduces an experiment conducted by lab 1 to evaluate the performance of a grasp planner); adapted from [2,18].
- **Replicated** experiments are conducted in a different lab, using the same context parameters and a different robot system configuration to **compare** the performance of multiple CUEs (e.g., comparing the performance of two grasp planners); adapted from [2,18].

Another commonly used term is "repeat"; while the proposed framework does not use this term explicitly, experiments that are conducted in the same lab, using the same context parameters and with the same robot system configuration (e.g., conducting multiple grasping trials on the same object to achieve statistically significant results) are **repeated** experiments (adapted from [2,18]). It should be noted that "reproducibility" and "replicability" are sometimes swapped or used interchangeably in the research literature, but the proposed conceptual framework explicitly differentiates them, following suit with the National Information Standards Organization (NISO) [2] and the Joint Committee for Guides in Metrology (JCGM) [18].

A diagram of the framework that relates the concepts of context, generalizability, reproducibility, comparison, and iterated, reproduced, and replicated experiments using an example scenario can be seen in Fig. 1. In this example representation of the framework, three labs each conduct two experiments:

- Lab 1 conducts experiment A_1 and **iterates** to conduct experiment B_1 with different target objects
- Lab 2 **reproduces** experiment A_1 to conduct experiment A_2 and **iterates** to conduct experiment C_2 with different lighting conditions

– Lab 3 **replicates** experiment A_1 to conduct experiment D_3 with a different grasp planner (CUE_D) and **iterates** to conduct experiment A_3 using the original grasp planner (CUE_A) simultaneously **reproducing** experiment A_1
– Across these six experiments, the following evaluations can be conducted:
  • **Generalizability** of CUE_A can be evaluated 7 times
  • **Reproducibility** of CUE_A can be evaluated 3 times
  • **Comparing** the performance of CUE_A and CUE_D 3 times

While a reproduced experiment may use the same context parameters and robot system configuration, there will be natural variations due to the complexities of real-world environments and nuances of physical systems. These natural variations have been demonstrated to substantially impact reproduced experiment results; for example, a 20% variation in performance was observed in [13] when reproducing an experiment across two labs. The authors suggest that "building precisely reproducible robotic setups is impossible and therefore absolute performance numbers on a benchmark task are meaningless." They instead
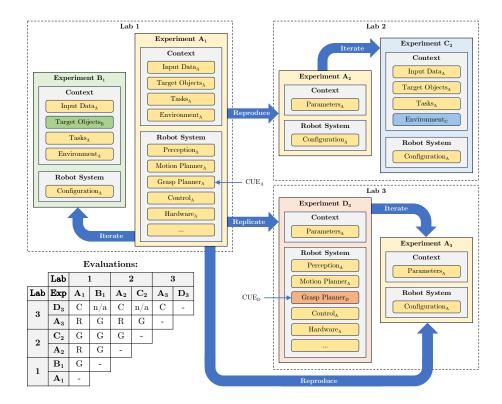

Evaluations:

Lab	Exp	1		2		3	
		A_1	B_1	A_2	C_2	A_3	D_3
3	D_3	C	n/a	C	n/a	C	-
	A_3	R	G	R	G	-	
2	C_2	G	G	G	-		
	A_2	R	G	-			
1	B_1	G	-				
	A_1	-					

Fig. 1. The MIRRER framework using an example scenario involving six experiments across three labs to evaluate the performance of two CUEs (grasp planners). The table in the bottom left shows the evaluations that can be performed across experiments (G = generalizability, R = reproducibility, C = comparison, n/a = not applicable).

recommend using a **local relative ranking (LRR)** evaluation method where each lab establishes a baseline within their own lab for comparison when benchmarking the performance of a new capability [13]. So then, following this recommendation when benchmarking to compare the performance of different CUEs, one should conduct both a **reproduced** experiment (to produce a local baseline with the original CUE) and a **replicated** experiment (to produce performance results of the new CUE). For example, using the scenario in Fig. 1, A_3 produces a local baseline of CUE_A in lab 3 that can then be compared to the results of D_3 which uses CUE_D to generate LRR_3. It should be noted that the LRR method was developed with the evaluation of software components in mind; producing a local baseline when the CUE is a piece of hardware such as a custom gripper (e.g., Yale OpenHand [24]) will introduce additional challenges (e.g., costs to purchase, fabrication expertise, logistics to share hardware between labs).

The authors of [13] also propose a method to globally rank several LRRs contributed by different labs. For example, if another lab were added to the scenario in Fig. 1 (lab 4) and reproduced experiments A_3 and D_3 from lab (i.e., A_4 and D_4), then LRR_3 and LRR_4 can be contributed towards a global ranking of the performance of CUE_A and CUE_D. Broad adoption of this method could significantly improve (and sufficiently limit) the interpretation of benchmarking results and highlight the need for reproduced/replicated experimentation.

3   Discussion

There are several factors to consider and gaps in existing technology and infrastructure before MIRRER experimentation can be conducted effectively.

Context characterization. While we define four high-level categories of context parameters that can influence the outcomes of an experiment, we have not specified the manner in which each parameter is recorded. We aim to strike a balance between ensuring that salient information is recorded and done so in a manner that is digestible, or "sufficiently complete," as the act of recording context should not become too cumbersome and risk discouraging researchers. There are some existing efforts in the research literature we can leverage to this end (e.g., ASTM standards F3218-19 [3] for environment characterization and F3381-19 [4] for describing objects). More work is needed to develop "sufficiently complete" context characterization methods, aiming to outline the unique parameters of context that have impacts on performance and can be controlled during experimentation.

Robot system configuration. All components of the robot system (especially the CUE) and the interoperability between components must be effectively specified such that it can be reproduced. On the hardware side, the make and model of robot arms, grippers, and sensors can be recorded, leveraging existing methods like ASTM F3327-23 [5] for characterizing robot configuration, but characterizing custom hardware will require more details or pointers to documentation. The use of standard bolt patterns, common mounts, and data connection methods (e.g., ethernet, USB, input/output terminals) simplifies the reproduc-

tion of physical connections between components, however publications are not always consistent when reporting robot system configuration [27]. On the software side, while enabling technologies like the Robot Operating System (ROS) have become ubiquitous for developing, executing, and connecting components, the field lacks consensus on software structure to ensure compatibility [27]. The development of new standards to specify common structures similar to IEEE 1873-2015 [17] (data formats for navigation maps) can support this issue.

Conducting experiments. Application of the MIRRER framework relies on researchers' ability to conduct multiple comparable experiments across different labs. At the very least, this will involve two experiments per lab (iterated experiments) in order to produce a LRR. While some component-level benchmarking methods are available (e.g., benchmarking motion planners [22]), they largely rely on simulation whereas physical testing requires a full robot system where holistic evaluations are typically conducted. A replicated experiment assumes that all components other than the CUE are consistent across labs, meaning that no changes are made to the robot system to accommodate the new CUE (e.g., modifying how data is passed between components to accommodate a particular execution pipeline). A current limitation in the field, though, is the lack of truly modular software components [27]; the existence of such would not only ease conducting replicated experiments, but would also mitigate the risk of robot system components other than the CUE significantly impacting performance results. An example of such a software pipeline is GRASPA [9] which has been used to compare the performance of multiple grasp planners. Such infrastructure as well as the new software standards mentioned previously would also lower the barrier to entry for research labs, both for benchmarking purposes and general utility of open-source software components.

Data storage and metrics. Effective use of MIRRER throughout the research field requires that results from evaluations be shared and stored in an accessible manner. Computer vision and machine learning communities regularly utilize assets like the Papers With Code repository [1], which links benchmarks from contributed papers with submitted scores from users and displays them in a leaderboard. In those domains, it is much simpler for one to recreate a submitted score using the submitted solution; in most cases, only a computer with the dataset used for benchmarking is required. As discussed previously, additional guidelines and tools are needed in order to enable this capability in robotics. Consideration must be given to the size of data that is stored, too, as the difference between reporting a series of performance metrics vs. an entire ROS bag of experiment data is significant. Similar to the distillation of relevant contextual parameters for characterizing an experiment, the same must be done to determine "sufficiently complete" elements that shall be recorded such that measures can be proven and effectively reproduced. Lastly, the metrics of generalizability and reproducibility are not formally defined and will require additional development. As a first step, simply reporting the resulting performance metrics from several iterated and reproduced experiments (respectively) and calculating the variance between them can represent either of these metrics.

Incentives. Despite a general consensus that improving reproducibility and replicability of robotics research is a worthy endeavor, the field does not yet provide sufficient incentives for researchers to do so, such as publication review criteria favoring research that includes comparison benchmarking [27] (which has become the norm for computer vision research). The human-robot interaction (HRI) domain does provide opportunities for researchers to publish this type of research by soliciting papers on replication studies at the HRI 2024 conference [16]. With performance results dependent on a human-in-the-loop, HRI research involves even more variables than what the MIRRER framework is scoped for, but there are no similar venues yet that provide value incentives for reproducing or replicating non-HRI research.

4   Conclusion and Future Work

This paper presents a conceptual framework, MIRRER, as a first step towards formalizing how to conduct evaluations of robot systems and their components in terms of generalizability, reproducibility, and replicability. We intend to continue development of the framework as well as hardware and software tools to enable application of MIRRER for experimentation. Several experiments have already been conducted within our own lab and with collaborating labs, with more planned now that the initial framework has been defined.

References

1. The Latest in Machine Learning - Papers With Code. `https://paperswithcode.com/`, accessed: 2024-04-28
2. Association for Computing Machinery (ACM): Artifact Review and Badging, `https://www.acm.org/publications/policies/artifact-review-and-badging-current`, accessed February 26, 2024
3. ASTM International: ASTM F3218-19 Standard Practice for Documenting Environmental Conditions for Utilization with A-UGV Test Methods. ASTM Book of Standards **15.13** (2019)
4. ASTM International: ASTM F3381-19 Standard Practice for Describing Stationary Obstacles Utilized within A-UGV Test Methods. ASTM Book of Standards **15.13** (2019)
5. ASTM International: ASTM F3327-23 Standard Practice for Recording the A-UGV Test Configuration. ASTM Book of Standards **15.13** (2023)
6. Bekiroglu, Y., Marturi, N., Roa, M.A., Adjigble, K.J.M., Pardi, T., Grimm, C., Balasubramanian, R., Hang, K., Stolkin, R.: Benchmarking protocol for grasp planning algorithms. IEEE Robotics and Automation Letters **5**(2), 315–322 (2019)
7. Bonsignorio, F.: A new kind of article for reproducible research in intelligent robotics [from the field]. IEEE Robotics & Automation Magazine **24**(3), 178–182 (2017)
8. Bonsignorio, F., Del Pobil, A.P.: Toward replicable and measurable robotics research [from the guest editors]. IEEE Robotics & Automation Magazine **22**(3), 32–35 (2015)

9. Bottarel, F., Altobelli, A., Pattacini, U., Natale, L.: Graspa-fying the panda: Easily deployable, fully reproducible benchmarking of grasp planning algorithms. IEEE Robotics & Automation Magazine (2023)
10. Burgess-Limerick, B., Lehnert, C., Leitner, J., Corke, P.: Dgbench: An open-source, reproducible benchmark for dynamic grasping. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3218–3224. IEEE (2022)
11. Cervera, E.: Try to start it! the challenge of reusing code in robotics research. IEEE Robotics and Automation Letters **4**(1), 49–56 (2018)
12. Cervera, E.: Run to the source: The effective reproducibility of robotics code repositories. IEEE Robotics & Automation Magazine (2023)
13. Dasari, S., Wang, J., Hong, J., Bahl, S., Lin, Y., Wang, A., Thankaraj, A., Chahal, K., Calli, B., Gupta, S., et al.: Rb2: Robotic manipulation benchmarking with a twist. arXiv preprint arXiv:2203.08098 (2022)
14. Dufrene, K., Nave, K., Campbell, J., Balasubramanian, R., Grimm, C.: The grasp reset mechanism: An automated apparatus for conducting grasping trials. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE (2024)
15. Gunes, H., Broz, F., Crawford, C.S., der Pütten, A.R.v., Strait, M., Riek, L.: Reproducibility in human-robot interaction: furthering the science of hri. Current Robotics Reports **3**(4), 281–292 (2022)
16. Human-Robot Interaction Conference: HRI 2024 Short Contributions, `https://humanrobotinteraction.org/2024/short-contribution/`, accessed March 7, 2024
17. IEEE Standards Association: IEEE 1873-2015 IEEE Standard for Robot Map Data Representation for Navigation, `https://standards.ieee.org/ieee/1873/5355/`, accessed March 7, 2024
18. Joint Committee for Guides in Metrology: International vocabulary of metrology—basic and general concepts and associated terms (VIM). JCGM **200**, 2012 (2012)
19. Khargonkar, N., Allu, S.H., Lu, Y., Prabhakaran, B., Xiang, Y., et al.: Scenereplica: Benchmarking real-world robot manipulation by creating reproducible scenes. arXiv preprint arXiv:2306.15620 (2023)
20. Leichtmann, B., Nitsch, V., Mara, M.: Crisis ahead? why human-robot interaction user studies may have replicability problems and directions for improvement. Frontiers in Robotics and AI **9**, 838116 (2022)
21. Leitner, J., Tow, A.W., Sünderhauf, N., Dean, J.E., Durham, J.W., Cooper, M., Eich, M., Lehnert, C., Mangels, R., McCool, C., et al.: The acrv picking benchmark: A robotic shelf picking benchmark to foster reproducible research. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 4705–4712. IEEE (2017)
22. Liu, S., Liu, P.: Benchmarking and optimization of robot motion planning with motion planning pipeline. The International Journal of Advanced Manufacturing Technology pp. 1–13 (2022)
23. Lynnerup, N.A., Nolling, L., Hasle, R., Hallam, J.: A survey on reproducibility by evaluating deep reinforcement learning algorithms on real-world robots. In: Conference on Robot Learning. pp. 466–489. PMLR (2020)
24. Ma, R., Dollar, A.: Yale openhand project: Optimizing open-source hand designs for ease of fabrication and adoption. IEEE Robotics & Automation Magazine **24**(1), 32–40 (2017)
25. Morrison, D., Corke, P., Leitner, J.: Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. IEEE Robotics and Automation Letters **5**(3), 4368–4375 (2020)

26. Norton, A., Saretsky, A., Yanco, H.: Developing metrics and evaluation methods for assessing ai-enabled robots in manufacturing. In: Proceedings of the AAAI spring symposium on artificial intelligence and manufacturing (2020)
27. Yanco, H., Norton, A., Calli, B., Dollar, A.: Collaborative Open-source Manipulation and Perception Assets for Robotics Ecosystem (COMPARE), `https://www.robot-manipulation.org/`, accessed February 27, 2024