Benchmarking Cluttered Robot Pick-and-Place Manipulation with the Box and Blocks Test

Andrew S. Morgan[1], *Student Member, IEEE,* Kaiyu Hang[1], *Member, IEEE,*
Walter G. Bircher[1], *Student Member, IEEE*, Fadi M. Alladkani[2], Abhinav Gandhi[2],
Berk Calli[2], *Member, IEEE,* and Aaron M. Dollar[1], *Senior Member, IEEE*

***Abstract*—In this work, we propose a pick-and-place benchmark to assess the manipulation capabilities of a robotic system. The benchmark is based on the Box and Blocks Test (BBT), a task utilized for decades by the rehabilitation community to assess unilateral gross manual dexterity in humans. We propose three robot benchmarking protocols in this work that hold true to the spirit of the original clinical tests—the Modified-BBT, the Targeted-BBT, and the Standard-BBT. These protocols can be implemented by the greater robotics research community, as the physical BBT setup has been widely distributed with the Yale-CMU-Berkeley (YCB) Object and Model Set. Difficulty of the three protocols increase sequentially, adding a new performance component at each level, and therefore aiming to assess various aspects of the system separately. Clinical task-time norms are summarized for able-bodied human participants. We provide baselines for all three protocols with off-the-shelf planning and perception algorithms on a Barrett WAM and a Franka Emika Panda manipulator, and compare results with human performance.**

I. Introduction

Enabling robots to work within, perceive, and manipulate their unstructured, human-made environment has motivated many decades of robotics research [1], [2]. Despite this longstanding research effort, there continues to be a vast ability gap between the tasks robots and humans are able to accomplish. This fact is perhaps most evident in the various robotic challenges within recent years, like the Amazon Picking Challenge (APC) [3], the DARPA Autonomous Robotic Manipulation (ARM) challenge [4], the Robot Grasping and Manipulation Competition 2016 [5], and the RoboCup@Home challenge [6], in which the robots can only demonstrate a tiny fraction of human dexterity, and require orders of magnitude more time to complete the same task.

The robotics community is lacking the tools to assess the manipulation performance of a given system and draw meaningful comparisons, which prevents systematic analysis, and therefore progress in the field. Unlike research disciplines that can be primarily evaluated by data sets and simulations (e.g. algorithms in image segmentation [7], 3D object retrieval [8], [9], object recognition [10], and SLAM [11]), robotic manipulation requires real-life experiments with physical objects and environments due to the difficulty of accurately simulating the contact phenomena. Nonetheless, in end-to-end task-oriented evaluations, such as those in the aforementioned challenges, it is difficult to evaluate individual components of the system (e.g. object recognition, object segmentation, motion planning, hardware design) and determine which component contributed to the success or failure of the task [12]. This is due to the holistic assessment and scoring of the performance with a high complexity task.

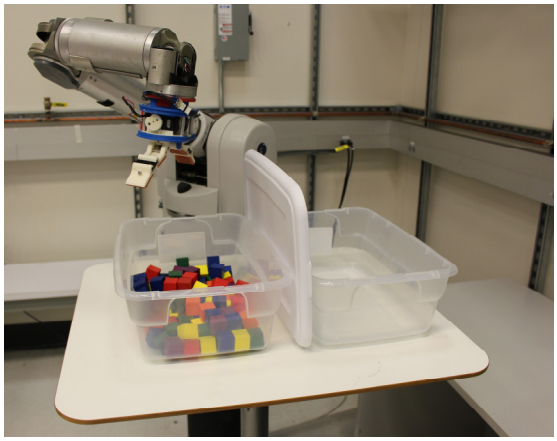

Fig. 1. Box and Blocks Test setup in front of a Barret WAM with a Yale Openhand Model T42. The goal is to transfer all 100 blocks from the start container to the goal container over the vertical obstacle.

In this work, we address this dearth of system evaluation in a standardized benchmark with protocols of increasing difficulty. By incrementally testing and challenging an additional system component between each protocol, namely, manipulator design, control, perception, and planning, we are able to evaluate various aspects of the system separately to a much greater extent and provide a general discussion of system limitations. Through this assessment, we enable investigators to objectively compare results for a more enlightened research discussion.

Our work defines three benchmarking protocols that are inspired by the clinical Box and Blocks Tests (BBTs). The BBT has been long utilized by clinicians in the rehabilitation community for evaluating upper-limb gross manual dexterity in physically impaired individuals. As originally standardized by Mathiowetz in 1985 [13], the standard test consists of two containers separated by a vertical barrier, with one container holding 150 colored wooden blocks. Within one minute, the participant must transfer as many single blocks as possible from the filled container to the empty container, ensuring that

Manuscript received: August, 13, 2019; Revised November, 18, 2019; Accepted December, 16, 2019.

This paper was recommended for publication by Editor Han Ding upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by the United States National Science Foundation under Grant IIS-1734190.

[1] A. S. Morgan, K. Hang, W. G. Bircher, and A. M. Dollar are with the Department of Mechanical Engineering and Materials Science, Yale University, USA ({Andrew.Morgan, Kaiyu.Hang, Walter.Bircher, Aaron.Dollar}@yale.edu).

[2] F. M. Alladkani, A. Gandhi, and B. Calli are with the Robotics Engineering Program, Worcester Polytechnic Institute, MA, USA ({fmalladkani, agandhi2, bcalli}@wpi.edu).

Digital Object Identifier (DOI): see top of this page.

TABLE I. STANDARD BOX AND BLOCKS TEST NORMS FOR HEALTHY INDIVIDUALS [13], [15]

Age Range	# Male	# Female	Male Avg.	Female Avg.	Weighted Avg.
6-7	26	33	52.55	56.05	54.51
8-9	30	32	61.75	61.60	61.67
10-11	43	40	67.15	68.80	67.95
12-13	34	36	73.50	72.05	72.75
14-15	34	34	75.60	73.75	74.68
16-17	31	35	78.95	75.65	77.20
18-19	33	30	79.55	76.95	78.31
20-24	29	26	87.30	85.70	86.54
25-29	27	27	84.55	83.45	84.00
30-34	27	26	81.60	82.70	82.14
35-39	25	25	80.85	84.15	82.50
40-44	26	31	81.50	80.40	80.90
45-49	28	25	76.35	80.20	78.17
50-54	25	25	78.00	76.00	77.00
55-59	21	25	74.50	74.15	74.31
60-64	24	25	70.90	74.85	72.92
65-69	27	28	67.90	71.65	69.81
70-74	26	29	65.30	68.45	66.96
75+	25	26	62.15	64.30	63.25

TABLE II. MODIFIED BOX AND BLOCKS TEST NORMS FOR 16 HEALTHY INDIVIDUALS (AGES 29.5 ± 8.9 YRS) [17]

Measure	Standing (s)	Sitting (s)
Right Hand Blocks	9.65 ± 0.86	9.70 ± 0.90
Left Hand Blocks	10.36 ± 1.12	10.38 ± 1.18

TABLE III. TARGETED BOX AND BLOCKS TEST NORMS FOR 19 HEALTHY INDIVIDUALS (AGES 29.9 ± 8.3 YRS) [18]

Measure	Standing (s)
Right Hand Blocks	25.8 ± 5.14

one's fingers cross the barrier. Due to the popularity of this test, normative data for healthy participants has been generated and is widely accepted for individuals in 19 age groups (ages 8-94). Additional works have validated the repeatability of the norms [14], extended the age groups from the original study [15], and even introduced added difficulties in order to asses precision dexterity (the modified box and blocks test [16], [17] and the targeted box and blocks test [18]). Able-bodied norms from these clinical variations provide inspiration and baseline comparisons for our proposed benchmarking protocols.

In robotic benchmarks, objects and environments should be standardized to ensure the experiments are conducted in comparable conditions. We rigorously describe the experimental setup, provide step-by-step instructions and evaluation metrics, and use the objects from Yale-CMU-Berkeley (YCB) Object and Model Set [12], which is widely available to the robotics research community. The proposed benchmarks follow very closely to the clinical BBTs, but differ in three ways as to conform to objects provided by the YCB Object Set: container geometries differ, the standard BBT only has 100 blocks, and the block template is smaller in order to fit into the bottom of the YCB container. All protocols are otherwise identical to their clinical counterpart.

II. RELATED WORK

In this section, we present BBTs in the literature, discuss the role of pick-and-place tasks in the literature, and give a summary of data sets and benchmarking efforts in the related field.

A. The Box and Blocks Test

The (standard) BBT has been utilized for evaluating upper limb manual dexterity in physically impaired individuals for several decades. Other tests have been proposed to evaluate similar measures, such as the Southampton Hand Assessment Procedure (SHAP) test [19], but are often more difficult to administer and can be expensive to purchase. As originally popularized by [13], the BBT study provided norms for able-bodied individuals, ages 20-92 in 12 age groups (318 females and 310 males). Recorded metrics signified the number of blocks transferred within one minute of testing per individual, while distinguishing by age group, sex, and hand of dominance. Follow up work added norms for 7 additional age groups from able bodied individuals ages 6-19 (231 females, 240 males), distinguishing by the same characteristics [15]. As to allow for a more meaningful comparison to these norms in robotics, weighted averages have been calculated for each age group by combining gender (female/male) and hand of dominance (right/left) categories from these two studies, presented in Table I, and represent the number of blocks transferred by each group within one minute.

Two altered box and blocks tests have been introduced in the literature, the modified BBT (2012) and the targeted BBT (2017), to better examine the kinematic repeatability of upper-limb trajectories. Healthy participant, normative data has been previously gathered for both altered tests (Table II and Table III) for 16 participants and 19 participants, respectively [17], [18]. The modified BBT assessment evaluated left-hand and right-hand execution times for right hand-dominant participants while standing and sitting. The targeted BBT assessment evaluated right hand transfer times for right-hand dominant participants. As the initial pose of the objects are fixed, these tests challenge precision arm and hand control of the user.

B. The Pick-and-Place Task

The BBT evaluation proposed in this work is a pick-and-place task as defined by [20], a type of task that has historically been of high interest to the robotics community largely due to its culmination of various problems in robotics and its applications in the real world. Pick-and-place applications often appear in Activities of Daily Living (ADLs), or tasks that would be required for home-focused autonomous service robotics. Example activities are outlined in a recent survey of human object manipulation [21]. Moreover, this interest is further underscored by well-publicized robotics challenges, e.g. [4], [6], and is also of great interest to the e-commerce industry for automated sorting and order fulfillment [3].

Though there is great interest in this type of task, efficient implementations have yet to be developed. Executions often suffer from being magnitudes slower compared to that of a human, and are less successful in completing the task. This can be attributed to several things, e.g. a subsystem contributing to a bottleneck or an inefficient integration of the subsystem. Increasing computational efficiency and efficacy of the subsystems becomes of high interest for effective task completion. For example, previous works have investigated accelerating grasp synthesis [22], simplifying control [23],

Fig. 2. (Left) M-BBT and T-BBT–block templates are affixed to the bottom of each container and 16 colored blocks (4 colors) are placed according to the template. (Right) S-BBT–100 blocks are placed randomly inside of the start container with random color distributions.

TABLE IV. SUMMARY OF METRICS TO REPORT

Metric	Description
Score	(M-BBT) A point is awarded if the correct block in order is successfully transferred to the goal container (T-BBT) A point is awarded if the correct block is successfully transferred to the goal container in the correct order and in the correct target location (S-BBT) A point is awarded for every transfer that consists of one or more blocks
Blocks Picked	(S-BBT only) Total number of blocks transferred to the target bin, greater than 'score' with picks of more than one block
Pick Attempts	(S-BBT only) Number of times the end effector tried to pick a block
End Effector Distance (m)	Distance traveled by the end effector during execution
Planning Time (s)	Amount of time used in motion and grasp planning
Execution Time (s)	Amount of time the manipulator or end effector is in motion
Total Time (s)	Amount of time used to complete the entire task

[24], and even using suction cups or specifically tailored manipulators to speed up the task [3], [25]. The BBT benchmarks in this work attempt to increasingly challenge each of these potential bottlenecks, promoting enlightened discussion and standardized evaluation for comparison.

C. Data Sets and Benchmarking

Object and model sets for benchmarking have been proposed in various forms and of differing scopes in the literature, e.g. [3], [26]–[29]. However, the sets often lack critical information required to carry out accurate simulations – such as object textures, 3D object models, object inertial properties, or coefficients of friction. Due to the complexity of distributing physical objects, one project has attempted to make a shopping list of objects for researchers to purchase, but this list is currently outdated [28]. Objects for the APC are available for purchase, but there remains an added barrier to accurately setting up the test environment.

Benchmarks in other applications that do not require physical objects have been significantly more successful, such as the creation of Imagenet [10] and the Princeton Shape Benchmark [9]. These datasets have become very large from collaboration with their associated communities. At its inception in 2009, Imagenet contained 3.2 million images and has since grown to over 14 million in just a decade. In few cases, benchmarking in physical systems has been proposed like that in Simultaneous Localization and Mapping (SLAM). In [30], a benchmark was proposed for indoor SLAM with physical robots by standardizing the environment and incorporating a reference robot for comparison between algorithms. Unlike these aforementioned benchmarks, there remains great merit in executing tasks in a physical environment, as execution in a simulation typically lacks reciprocity to the real world where robots much surely work [31]. For this reason, we select the use of the YCB set, an invaluable tool for creating physical benchmarking protocols for the robotics community, as the object set has been distributed to over 120 research groups at the time of writing. Previously, an end-to-end benchmark using this object set has been proposed for assessing the picking performance of a robot from a standard shelf [32].

III. BENCHMARKING PROTOCOLS

Three clinically-inspired benchmarking protocols based on the modified, targeted, and standard BBTs are described in this section (full instructions of the experimental procedure and scoring criteria are provided as a multimedia attachment). In the first benchmark, the Modified-BBT (M-BBT) [16], the goal is for the robot to transfer 16 identically oriented blocks from one container to the other over a separating barrier (container lid) in minimal time. This task mainly challenges manipulator design and grasping. The second protocol, the Targeted-BBT (T-BBT) [18], begins similarly to the M-BBT but requires precision placement of the block on the other side of the barrier. This task further challenges the accuracy and control of the end effector and of the manipulator to ensure the dynamics associated with object placement do not incur undesirable object movement upon release. The third and final protocol mimics that of the Standard-BBT (S-BBT) [13], where the task is to transfer as many randomly configured blocks as possible (out of 100) across the barrier in minimal time. There are two variations of this third protocol, the first is timed for one minute, as to allow for comparison to the clinical evaluation, and the second is untimed. This final protocol further challenges perception and planning, as object segmentation and grasp synthesis in cluttered environments remain difficult problems in robotics.

The physical setup of all non-manipulated objects is identical for all three protocols (YCB Obj. #68, 69) [12]. Start

and goal containers are positioned on a support surface in front of the manipulator, close enough that the entire volume of the bin is reachable. Relative location of the containers in front of the robot is left up to the user, but the containers and the barrier lid must be within the same relative configuration as depicted in Fig. 2. That is, the two containers are pushed together length-wise with one container's lid acting as a separating barrier. Blocks (YCB Obj. #70) [12] are placed inside of one of the containers and are oriented according to the specific benchmarking protocol being tested. The determination of which bin is filled (start container) and which bin is empty (goal container), is left up to the user. For all protocols, the end effector must start in a position outside of either container.

An overarching goal of the proposed benchmarks is to evaluate all potential implementations for general pick-and-place tasks with a standardized test. Therefore, hardware and software implementations used in execution are not restricted in any way—types of manipulators, end effectors, and sensing modalities are free to be determined by the user. However, alteration of physical objects provided in the YCB set is prohibited. This includes a restriction on changing colors, textures, or weights of the blocks or containers. Additionally, markers cannot be placed on any of the blocks, but may be placed on the container for pose recognition. Container position and orientation can either be determined *a priori* or during execution. The bins must remain in the same configuration during the entirety of the task and cannot be moved purposefully for object reorientation. If the center of the bin moves more than $2.54cm$, or the width of a cube, from the original starting position, or the bin rotates more than $10°$ about the center of the container, the task must be restarted and the score for that task execution is zero.

Two mirrored templates, one for the left container and one for the right container, are provided for the M-BBT and the T-BBT protocols (provided as a multimedia attachment). Each template has sixteen $3.2cm\ x\ 3.2cm$ numbered target block locations oriented in groups of four in a row (Fig. 2). Rows are separated by $2cm$ from one another. The template is enclosed by a rounded rectangle mimicking the shape of the container's bottom. Templates are affixed to the bottom of the container during execution. To ensure the template is appropriately placed, block 1 should be the outermost and furthest block target location from the manipulator.

Specific scoring rules differ between all three protocols. In general, for Protocols I and III that do not require precision object placement, a successful transfer is characterized similarly to [13] and requires that the object fully reaches onto the other side of the barrier before dropping into the goal container, i.e. blocks cannot be thrown over the barrier but can be released from any elevation above the goal container. Scores are not penalized if blocks bounce out of the goal container after release, but still count as a single point. These rules are in-line with the clinical protocol.

Each task protocol should be completed consecutively at a minimum of five times as to allow for a general understanding of the system's robustness. Failed tasks, as defined individually for each protocol, result in a score of zero for that execution. In addition to the score for each protocol, the amount of time in seconds used in planning, in execution, and in total needs to be reported. Time dedicated to perception and decision making should constitute the difference of the total time with execution time and planning time. The total distance that the end-effector traveled in meters during execution (most distal link of the manipulator) must also be recorded. A summary of reported metrics is provided in Table IV.

A. Protocol I: Modified - Box and Blocks Test (M-BBT)

The first protocol, the Modified-BBT (M-BBT) [16], simplifies the perception, planning, and control problem to focus on manipulator design and execution speed. A total of sixteen colored blocks, consisting of four different colors determined by the user, are placed according to the provided template in either the left or right container. Blocks must start inside of the designated starting locations and blocks of the same color must be placed in the same row (Fig. 2)

The goal for this evaluation is to move all sixteen blocks from one bin to the other in the correct order and in minimal time. Blocks must be transferred one at a time, starting with block 1 and ending with block 16. Blocks do not have a target location inside of the goal container. If neighboring blocks are perturbed by the end effector during execution, the task can be continued as long as the same order of blocks are picked as defined by the beginning of the task. If a pick is missed, the system must continue to the next block. In cases where the start container moves more than $2.54cm$, two blocks are picked at once, or the wrong picking order is executed, the execution receives a score of 0. The maximum score for this protocol is 16 and occurs when all blocks are transferred in the correct order.

B. Protocol II: Targeted - Box and Blocks Test (T-BBT)

The second protocol, the Targeted-BBT (T-BBT) [18], builds off of the M-BBT as to require dynamic placement control of the object, further challenging the control of the manipulator. The task environment is setup similarly to Protocol I, but now requires that each block is placed within a specific target location. In minimal time the goal is to transfer each block, in order from 1-16, from the start container to the goal container by matching the pick location number with the place location number, and within the $3.2cm\ x\ 3.2cm$ target location. If neighboring blocks during the pick are knocked by the end effector during execution, the task can be continued as long as blocks are picked in the same order as defined by the beginning of the task. If a pick is missed, the system must continue to the next block, and therefore the target location for that block in the target bin should remain empty.

Block placement and control becomes pivotal for scoring points. Once a block is picked, the manipulator must complete placement of that block before moving on to the next pick. The block can either be directly placed by the end effector or can be placed into the goal container and slid into the desired position using environmental affordances. Once another block is picked, the user can no longer manipulate already placed blocks. If other blocks are perturbed during placement of a single block, the pose of those perturbed blocks cannot be deliberately changed afterwards. Points are awarded at the end of the task, and signify that a corresponding start location and goal location match and that the block is completely inside of the target location. As before, in cases where the start container moves more than $2.54cm$, two blocks are picked at once, or the wrong picking order is executed, the task receives a score

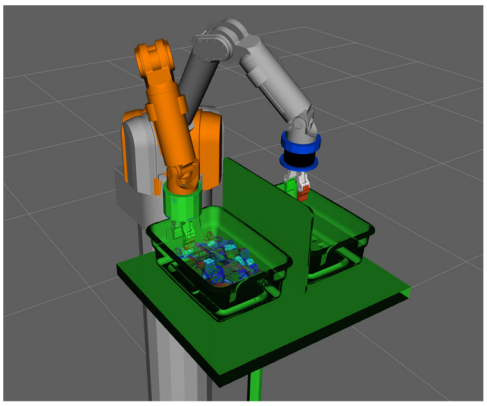

Fig. 3. Geometric constraints used for motion planning on a Barrett WAM manipulator. The gray arm (right) is the current state and the green arm (left) is evaluating collisions. Blocks inside of the container are visualized as a point cloud.

of 0. The maximum score for this protocol is 16 and occurs when all blocks were placed inside of the target locations in the correct order.

C. *Protocol III: Standard - Box and Blocks Test (S-BBT)*

The third protocol, the Standard-BBT (S-BBT) has two parts and presents additional difficulties in perception, since objects must be segmented in an occluded environment, and motion/grasp planning, as the manipulator must work in a cluttered environment. This protocol requires that one bin, the start container, is filled with all 100 colored blocks of an even color distribution throughout the container. This can be achieved by placing the lid on the container after filling and shaking the container. Inside of the container, four blocks of the same color should not be adjacently touching one another. Moreover, due to the size of the container and the total number of blocks used in this protocol, there should not exist a stack of blocks in the box that is more than two objects high.

This protocol has two tasks, the first is timed for one minute (Protocol IIIa) and the second is untimed (Protocol IIIb). In both tasks, the goal is for the user to transfer as many blocks as possible across the barrier and into the goal container in minimal time. Picks of more than one block only count as a single point. The user may find strategy in initially transferring more than one block in the beginning of the task to mitigate clutter, but this also limits the total number of points the user can score at the end of the task. In the one-minute timed task, if the manipulator has a successful pick once time expires, a transfer can be recorded as a point.

The untimed task, Protocol IIIb, presents interesting problems in motion planning, grasp planning, and control. Once the top layer of blocks is removed, the remaining layer typically lacks gaps in which for finger insertion (Fig. 6). Once a finger is inserted and a grasp is acquired, the planner must then account for other blocks in the bin, as collision with these objects will likely perturb the container undesirably. The end of the task is determined by the user, which likely occurs when the planner can no longer plan a grasp or all blocks are removed from the container. As with all protocols, if the container moves beyond its allotted translational and rotational threshold, the test fails with a score of 0. The maximum score for this protocol is 100.

Fig. 4. Franka Emika Panda setup with adapted fingertips.

IV. Baseline Implementations

All three protocols were attempted with two different robotic systems using off-the-shelf planning and perception algorithms to determine baseline results. In the first setup, an underactuated Yale Openhand Model T42 [33] (pivot-flexure model) powered by two Dynamixel RX-28 actuators was affixed to a Barrett WAM manipulator. A support surface (60cm x 70cm) was placed 12cm directly in front of the manipulator. The BBT setup was placed in the center of the support surface (Fig. 1, 3). A Microsoft Kinect was mounted overhead providing a point cloud of the environment. Geometric collision constraints were configured and velocity control motion planning was achieved with a RRT-Connect planner [34] in a MoveIt! environment. Geometric container and barrier object models were created to define the collisions within the environment (provided as multimedia attachments). The point cloud was segmented such that only blocks inside of the filled container were available (Fig. 3). Block position estimation was achieved through the use of a KMeans++ algorithm subject to the location and color of the points. After each pick, the number of specified clusters was reduced by one and the object position estimations were recomputed. The system determined on which block to approach by either order (protocols 1 and 2) or height (protocols 3a and 3b). The motion planner then computed a trajectory to place the fingers directly above the desired block for grasping. Additional waypoints were added for precision pick, which were located directly above the block at increments of 5cm.

The second system was a Franka Emika Panda (Fig. 4). The standard gripper was modified such that its fingers were extended by 140 mm using 3D printed parts, since the gripper itself was too wide to fit into the container. Even though the printed parts provide some level of compliance to the gripper, it is still quite rigid compared to the Model T42 gripper used in the first setup. Again, different from the first setup, an eye-in-hand system was used with an Intel Realsense D435i depth sensor mounted at the end effector, right above the gripper base. The two containers were placed 10cm in front of the robot. The point cloud data was transformed into the robot frame and is segmented using an off the shelf Euclidian cluster extraction algorithm from the Point Cloud Library. Centroids for each of the blocks were computed by averaging the point cloud data. MoveIt! was utilized and occlusions were defined

TABLE V. WAM BASELINE EXECUTION SCORES FOR ALL THREE BOX AND BLOCKS PROTOCOLS

	Score	Dist. (m)	Planning (s)	Execution (s)	Total (s)
	Protocol I: Modified BBT				
1	16	31.82	64.97	172.87	256.98
2	16	28.29	41.63	180.18	239.79
3	16	27.19	61.82	173.21	253.97
4	16	29.59	45.32	170.25	234.07
5	16	28.07	44.59	168.93	231.97
Avg	16.00	28.99	51.66	173.09	243.36
Std	0.00	1.61	9.71	3.89	10.26
	Protocol II: Targeted BBT				
1	2	42.3	41.27	221.95	274.21
2	2	44.12	59.89	224.47	295.37
3	4	44.42	48.26	217.6	227
4	1	43.32	45.47	221.86	279.81
5	0	43.57	52.48	222.81	283.24
Avg	1.80	43.54	49.47	221.74	281.90
Std	1.48	6.71	7.11	2.53	8.23

	Score	Blocks Picked	Pick Att.	Dist. (m)	Planning (s)	Execution (s)	Total (s)
	Protocol IIIa: Timed Standard BBT						
1	5	8	5	8.48	3.06	44.26	60
2	5	7	5	7.77	3.53	54.74	60
3	4	7	5	11.02	3.14	47.87	60
4	5	6	5	9.2	7.11	44.04	60
5	5	9	5	7.59	4.71	55.09	60
Avg	4.80	7.40	5.00	8.81	4.31	49.29	60.00
Std	0.40	1.02	0.00	1.24	1.52	4.86	0.00
	Protocol IIIb: Untimed Standard BBT						
1	36	61	52	100.97	159.12	629.81	1024.27
2	30	67	43	88.23	125	505.6	824.21
3	33	58	40	79.26	116.04	492.63	783.13
4	31	53	45	84.81	116.97	531.36	843.02
5	32	62	54	111.12	164.19	662.15	1084.5
Avg	32.40	60.20	46.80	92.99	136.26	564.31	911.83
Std	2.30	5.17	5.97	12.95	23.51	76.70	133.64

TABLE VI. PANDA BASELINE EXECUTION SCORES FOR ALL THREE BOX AND BLOCKS PROTOCOLS

	Score	Dist. (m)	Planning (s)	Execution (s)	Total (s)
	Protocol I: Modified BBT				
1	16	33.16	7.75	205.7	245.04
2	11	22.13	6.48	189.86	221.29
3	16	29.97	8.15	214.6	254.87
4	15	34.43	8.05	237.44	277.81
5	14	34.39	6.8	206.68	240.69
Avg	14.40	30.82	7.45	210.86	247.94
Std	2.07	5.18	0.75	17.36	20.68
	Protocol II: Targeted BBT				
1	14	50.9	9.3	301.2	340.5
2	13	51.4	10.2	286.1	327.9
3	12	39.8	8.3	240	274.2
4	13	36.7	9.5	267.4	306.3
5	13	41.4	8.2	251.5	286.7
Avg	13.00	44.00	9.10	269.31	307.10
Std	0.70	6.70	0.80	24.80	27.50

	Score	Blocks Picked	Pick Att.	Dist. (m)	Planning (s)	Execution (s)	Total (s)
	Protocol IIIa: Timed Standard BBT						
1	3	3	4	10.89	1.74	50.53	60
2	2	2	4	8.74	1.88	50.15	60
3	3	3	4	6.35	1.96	48.27	60
4	4	4	5	7.79	1.83	50.97	60
5	2	4	4	7.43	1.95	49.23	60
Avg	2.80	3.20	4.20	8.24	1.87	49.83	60.00
Std	0.80	0.83	0.40	1.71	0.09	1.08	0.00
	Protocol IIIb: Untimed Standard BBT						
1	9	9	14	28.9	6.83	176.15	215.06
2	11	11	33	68.165	15.57	475.39	556.8
3	6	6	14	20.48	13.7	193.02	233.12
4	6	6	11	26.65	8.2	254.83	296.14
5	6	6	7	15.538	3.96	114.25	166.49
Avg	7.60	7.60	8.60	31.95	9.65	248.73	293.52
Std	2.30	2.30	3.71	20.91	4.84	132.94	154.32

in the same way as the first setup. The trajectories were generated in Cartesian space to prevent undesired contact with the blocks other than the target block.

Each protocol was evaluated with five consecutive executions for both systems as presented in Table V and Table VI, and Fig. 5.

A. The Modified Box and Blocks Test Baseline

Protocol I was implemented on the specified setups. When using the WAM setup, all 16 blocks were successfully transferred over the barrier in the correct order in each of the 5 executions. In two executions, the fingers undesirably interacted with neighboring blocks, moving the neighbors less than $2cm$, but did not provide a large enough perturbation to affect the system. While the planning time varied significantly between executions with an average time of 51.66 ± 9.71s, the manipulator's execution time was similar for each trial. This variation in planning time can be attributed to the random search implemented in the motion planner.

During the Panda executions, similar undesired contacts occurred with the blocks neighboring the target block. In the second execution, the arm got into joint singularities and the execution terminated without attempting to pick all the blocks. This problem is due to trying to achieve a fast cartesian space planner, and not checking the joint constraints along the trajectory. Nevertheless, this setup was not as successful as the WAM-Model T42 setup for recovering from these situations due to the rigidity of the gripper. Here, we see the advantage of using compliant grippers in compensating for uncertainties during the picking operation.

B. The Targeted Box and Blocks Test Baseline

The targeted BBT requires the system to precisely place the blocks after each pick, considering the dynamics of block placement. In this test, the WAM setup scored very low compared to the first protocol (1.8 ± 1.48 blocks). This is mainly due to the lack of precision of the compliant hand on a low-impedance manipulator, which resulted in unpredictable

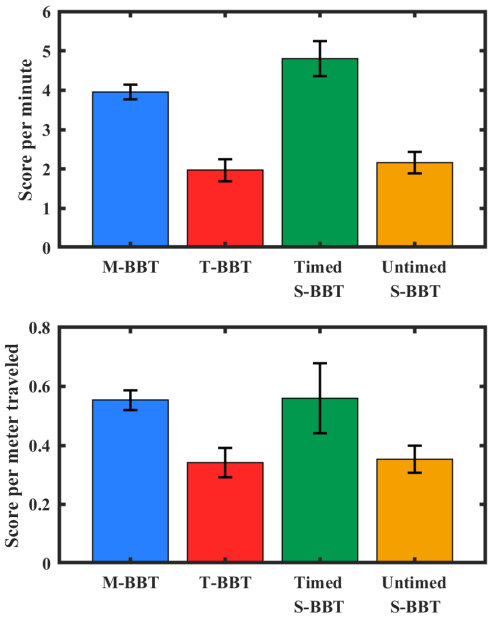

Fig. 5. (Top) Ratio of score to total task time for the WAM results. (Bottom) Ratio of score to end effector travel distance during execution for the WAM. Larger ratios typically signify higher task efficiency. Due to the nature of the task, the T-BBT and Untimed S-BBT require additional planning to ensure the block is picked (and placed) vertically to minimize interactions with neighboring blocks.

object motion during release especially if one finger lost contact before the other. Comparatively, the Panda setup scored much higher (13 ± 0.7 blocks), taking advantage of the precision evident in rigid manipulators and grippers. This outcome further underscores the compliance/rigidity tradeoff.

C. The Standard Box and Blocks Test Baseline

The final protocol was evaluated in both timed and untimed settings for each of the manipulators. Within the allotted minute of the timed task, the WAM setup was successfully able to transfer 4.8 ± 0.4 blocks on average. Two-block transfers were often executed in order to simplify planning. The average manipulator execution time (49.29 ± 4.86s) was over twelve times greater than that of the time required to create a plan (4.39 ± 1.52s). Though now adding the difficulty of perceiving individual blocks, the planning involved in this task was easier than others, as the end effector was able to interact with neighboring blocks without penalty, since all blocks were picked from the top layer. End effector distances also deviated between executions, with an average of 8.81 ± 1.24m. This again is attributed to the random joint configuration search implemented by the planner. The performance of the WAM setup was better than that of the Panda, due to the advantage of using a compliant system in cluttered, unstructured environments.

In the untimed evaluation, the performance difference between the WAM and Panda can be seen more clearly. Given enough time, the WAM setup picked significantly more blocks compared to the Panda setup due to compliance. In these tests, it was easily noted how difficult it was for the grippers to be accurately inserted into the bottom layer of the container while in a cluttered environment (Fig. 6). In the WAM setup, compliance was leveraged by devising an alternative strategy, which was to place the hand on top of a row of blocks and rotate the wrist before attempting to grasp. While rotating, the hand would continue to push into the container to insert the fingertip. This allowed the fingers to reconfigure the blocks before attempting to grasp.

Fig. 6. (Left) Inserting a finger into the bottom layer of bin is difficult as there are few gaps for insertion. (Right) Blocks around the perimeter of the left start container are not transferred during the untimed task as a grasp plan was difficult to find.

The untimed task on the WAM resulted in an average score of 32.4 ± 2.3 blocks picked over 911.8 ± 133.6s, far greater than what was achieved on the Panda (7.6 ± 2.3 blocks). Efficiency of the executions averaged 2.15 ± 0.27 blocks per minute, less than half that of the average of the timed test. This deviation can be attributed to the added planning and control difficulty with finger insertion into the bottom layer. Not all blocks were able to be picked out of the container, as only 60.2 ± 5.17 total blocks on average were transferred. Blocks not picked before termination were typically around the perimeter of the box, and a grasp plan was never found, as in Fig. 6.

V. Discussion

The executions presented in Sec. IV serve as baseline implementations for all three benchmarking protocols. In all executions, we recognize a noticeable difference between execution time and planning time, where the execution speed of the manipulator often contributed to more than 70% of the time used. While increasing the speed of the manipulator may contribute to a faster execution time, we noticed that it decreased the accuracy of both manipulators, presenting a bottleneck in their designs.

The execution time could have been decreased with an optimal trajectory planner, where the RRT-Connect architecture resulted in large end effector distance variations within the tasks. Off-the-shelf optimal trajectory planners were found to be too slow to use in execution. As in Fig. 5, the M-BBT and timed S-BBT resulted in the highest scores per minute, which is consistent with the clinical trials. This is largely due to the fact that the T-BBT required time to precisely place the object while avoiding interactions with neighboring blocks. Additionally, the untimed S-BBT was difficult due to the finger insertion problem. Both, the T-BBT and the untimed S-BTT, would have benefited from more advanced control for increased precision. Similar results are portrayed in the scores per meter traveled comparison, as the majority of time used was dedicated to manipulator movement. As faster manipulators and planning algorithms are used, scores, and consequently these ratios, will increase.

Due to similarities between the clinical and robot protocols, we are also able to generally compare task execution times between robots and humans. For example, in the WAM implementation with off-the-shelf components, the M-BBT was 24 times slower and the T-BBT was over 14 times slower than that of a human, from the age groups presented in [17], [18]. The timed S-BBT was over 11 times slower than that of a 6-7 year-old child, further underscoring the vast ability gap between a robot and a human in this task.

VI. Conclusion

In this work, we identify the inability to separate the components used in most task-level benchmarks and propose three standardized tasks based on the clinically utilized Box and Blocks Test. Benchmarking using the BBT is advantageous as not only has it been utilized for decades in the rehabilitation community to provide baseline able-bodied norms for comparison, but the physical setup is also included in the widely distributed YCB Object and Model set. Due to its significance in the rehabilitation community and its large distribution, it provides an accessible platform for evaluating the pick-and-place task.

Three protocols were designed by challenging an additional system component at each level. For each of the three benchmarks, we provide baseline results using off-the-shelf planning and perception algorithms on a Barrett WAM and a Franka Emika Panda. We compare baseline results to human performance, finding that robot execution times are over ten slower. By evaluating these benchmarks with different manipulators, planners, control algorithms, and perception systems, the protocols provide objective measures for researchers in the robotics community to compare approaches. We recognize there is much work still to be completed in the field of manipulation and it is our hope that these tests provide insight for future evaluation towards human-level pick-and-place efficiency.

References

[1] R. M. Murray, Z. Li, and S. Sastry, *A mathematical introduction to robotic manipulation*. CRC Press, 1994.

[2] A. M. Okamura, N. Smaby, and M. R. Cutkosky, "An overview of dexterous manipulation," in 2000 *IEEE International Conference on Robotics and Automation.* vol. 1, pp. 255–262.

[3] N. Correll *et al.*, "Analysis and Observations From the First Amazon Picking Challenge," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 1, pp. 172–188, Jan. 2018.

[4] "Autonomous Robotic Manipulation (ARM)." [Online]. Available: https://www.darpa.mil/program/autonomous-robotic-manipulation.

[5] J. Falco, Y. Sun, and M. Roa, "Robotic Grasping and Manipulation Competition: Competitor Feedback and Lessons Learned," Springer, Cham, 2018, pp. 180–189.

[6] J. Stuckler, D. Holz, and S. Behnke, "RoboCup@Home: Demonstrating Everyday Manipulation Skills in RoboCup@Home," *IEEE Robot. Autom. Mag.*, vol. 19, no. 2, pp. 34–42, Jun. 2012.

[7] "The Object Segmentation Database(OSD)." [Online]. Available: https://www.acin.tuwien.ac.at/en/vision-for-robotics/software-tools/osd/.

[8] "Digital Shape Workbench - Shape Repository v5.0." [Online]. Available: http://visionair.ge.imati.cnr.it:8080/ontologies/shapes/releases.jsp.

[9] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The princeton shape benchmark," in *Proceedings Shape Modeling Applications, 2004.*, pp. 167–388.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

[11] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 573–580.

[12] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set," *IEEE Robot. Autom. Mag.*, vol. 22, no. 3, pp. 36–52, Sep. 2015.

[13] V. Mathiowetz, G. Volland, N. Kashman, and K. Weber, "Adult Norms for the Box and Block Test of Manual Dexterity," *Am. J. Occup. Ther.*, vol. 39, no. 6, pp. 386–391, Jun. 1985.

[14] J. Desrosiers, G. Bravo, R. Hébert, E. Dutil, and L. Mercier, "Validation of the Box and Block Test as a measure of dexterity of elderly people: reliability, validity, and norms studies.," *Arch. Phys. Med. Rehabil.*, vol. 75, no. 7, pp. 751–5, Jul. 1994.

[15] V. Mathiowetz, S. Federman, and D. Wiemer, "Box and Block Test of Manual Dexterity: Norms for 6–19 Year Olds," *Can. J. Occup. Ther.*, vol. 52, no. 5, pp. 241–245, Dec. 1985.

[16] J. S. Hebert and J. Lewicke, "Case report of modified Box and Blocks test with motion capture to measure prosthetic function.," *J. Rehabil. Res. Dev.*, vol. 49, no. 8, pp. 1163–74, 2012.

[17] J. S. Hebert, J. Lewicke, T. R. Williams, and A. H. Vette, "Normative data for modified Box and Blocks test measuring upper-limb function via motion capture," *J. Rehabil. Res. Dev.*, vol. 51, no. 6, pp. 918–932, 2014.

[18] K. Kontson, I. Marcus, B. Myklebust, and E. Civillico, "Targeted box and blocks test: Normative data and comparison to standard tests.," *PLoS One*, vol. 12, no. 5, p. e0177965, 2017.

[19] J. Adams, K. Hodges, J. Kujawa, and C. Metcalf, "Test-retest Reliability of the Southampton Hand Assessment Procedure," *Int. J. Rehabil. Res.*, vol. 32, p. S18, Aug. 2009.

[20] T. Lozano-Perez, J. L. Jones, E. Mazer, and P. A. O'Donnell, "Task-level planning of pick-and-place robot motions," *Computer (Long. Beach. Calif).*, vol. 22, no. 3, pp. 21–29, Mar. 1989.

[21] Y. Huang, M. Bianchi, M. Liarokapis, and Y. Sun, "Recent Data Sets on Object Manipulation: A Survey," *Big Data*, vol. 4, no. 4, pp. 197–216, Dec. 2016.

[22] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-Driven Grasp Synthesis—A Survey," *IEEE Trans. Robot.*, vol. 30, no. 2, pp. 289–309, Apr. 2014.

[23] N. Hogan, "Impedance Control: An Approach to Manipulation," in *1984 American Control Conference*, 1984, pp. 304–313.

[24] M. Gabiccini, A. Bicchi, D. Prattichizzo, and M. Malvezzi, "On the role of hand synergies in the optimal choice of grasping forces," *Auton. Robots*, vol. 31, no. 2–3, pp. 235–252, Oct. 2011.

[25] V. Nabat, M. de la O Rodriguez, O. Company, S. Krut, and F. Pierrot, "Par4: very high speed parallel robot for pick-and-place," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 553–558.

[26] N. Yozbatiran, L. Der-Yeghiaian, and S. C. Cramer, "A Standardized Approach to Performing the Action Research Arm Test," *Neurorehabil. Neural Repair*, vol. 22, no. 1, pp. 78–90, Jan. 2008.

[27] S. Kalsi-Ryan, A. Curt, M. C. Verrier, and M. G. Fehlings, "Development of the Graded Redefined Assessment of Strength, Sensibility and Prehension (GRASSP): reviewing measurement specific to the upper limb in tetraplegia," *J. Neurosurg. Spine*, vol. 17, no. Suppl1, pp. 65–76, Sep. 2012.

[28] Y. S. Choi, T. Deyle, T. Chen, J. D. Glass, and C. C. Kemp, "A list of household objects for robotic retrieval prioritized by people with ALS," in *2009 IEEE International Conference on Rehabilitation Robotics*, pp. 510–517.

[29] "Amazon Picking Challenge - The Future of Robotics." [Online]. Available: http://amazonpickingchallenge.org/.

[30] C. Sprunk *et al.*, "An Experimental Protocol for Benchmarking Robotic Indoor Navigation," Springer, Cham, 2016, pp. 487–504.

[31] R. A. Brooks and M. J. Mataric, "Real Robots, Real Learning Problems," in *Robot Learning*, Springer US, 1993, pp. 193–213.

[32] J. Leitner *et al.*, "The ACRV picking benchmark: A robotic shelf picking benchmark to foster reproducible research," in *2017 IEEE International Conference on Robotics and Automation*, pp. 4705–4712.

[33] R. Ma and A. Dollar, "Yale OpenHand Project: Optimizing Open-Source Hand Designs for Ease of Fabrication and Adoption," *IEEE Robot. Autom. Mag.*, vol. 24, no. 1, pp. 32–40, Mar. 2017.

[34] J. J. Kuffner and S. M. LaValle, "RRT-connect: An efficient approach to single-query path planning," in *2000 IEEE International Conference on Robotics and Automation.*, vol. 2, pp. 995–1001.