Qualitative Comparison between Marker-Based and Video-Based Human Pose Estimation in the context of Imitation Learning

Max Lödige[1⋆], Alexander Fabisch[2], and Lisa Gutzeit[1]

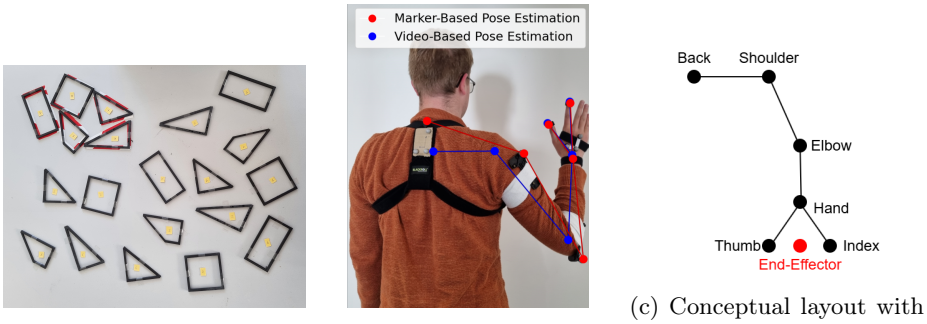
[1] University of Bremen, Robotics Research Group, Bremen, Germany
{loedige,Lisa.Gutzeit}@uni-bremen.de
[2] German Research Center for Artificial Intelligence (DFKI GmbH), Robotics Innovation Center, Bremen, Germany
Alexander.Fabisch@dfki.de

Abstract. Imitation learning from human demonstration often relies on marker-based motion capture, which restricts data recording to controlled environments. Video-based motion capture presents an alternative that reduces the recording effort. To compare both, we develop a novel evaluation framework using four categories of metrics, namely pose estimation quality, trajectory distance, behaviour separability, and grasp characterization, applied to three state-of-the-art video-based 3D pose estimation models in a pick-and-place scenario with parallel recording. Using this framework, we show that video-based pose estimation using a depth map captures actions and trajectories well enough for imitation learning, but high-fidelity movement such as grasping remains beyond the current capabilities of all evaluated video-based modalities.

1 Introduction and Related Work

As robots are increasingly used in highly dynamic environments, traditionally programmed behaviours reach their limit in adaptability and thus result in the development of more dynamic learning pipelines. Imitation learning is one approach for such a dynamic learning pipeline and is shown to work well in robotic learning paradigms due to its efficiency in teaching tasks that are hard to express as reward function in traditional reinforcement learning [7]. The subfield of imitation learning from human demonstration places a human in the role of the expert, in which a robot learns how to perform a task by observing the human. To gather such expert data, current approaches mainly rely on external tools such as marker-based motion capture or controllers [15]. These tools provide highly accurate data but restrict recording to controlled laboratory environments. A potential alternative is video-based human pose estimation, a field of computer vision that estimates human keypoint locations from RGB or RGBD data, often from a single camera. This drastically reduces the sensing requirements for a

⋆ Research conducted while affiliated to University of Bremen.



(a) Pick-and-place table. (b) Markers and keypoints. (c) Conceptual layout with virtual end-effector point.

Fig. 1: Experimental setup.

robotic system and enables imitation learning from human demonstration outside of carefully prepared laboratory settings, at the cost of reduced accuracy. This work's main contribution is the development of an evaluation framework designed to assess the quality of human pose estimation in the context of imitation learning. Traditional imitation learning evaluation measures the trained agent's behaviour (e.g., task success rate), which introduces confounding variables such as morphological differences or algorithmic bias. To evaluate the data quality independently from any single imitation learning technique, we introduce a novel framework comprising four metric categories: pose estimation quality, trajectory distance, behaviour separability, and grasp characterization. We utilize marker-based motion capture as the ground truth, as it is the recognised gold standard for imitation learning from human demonstration, and apply the framework to compare it against three distinct video-based motion capture approaches.

2 Study

Pick-and-place is a canonical manipulation primitive that underpins many household and assembly tasks, making it a representative testbed for imitation learning. The study design follows the diverse behaviours methodology of Jia et al. [13], applying the principles of diverse behaviour, multiple demonstrations, variable trajectory lengths, and task variations [13, Chapter 3.2]. The scenario consists of five uniquely shaped blocks, each sorted from one of three fixed starting positions to a fixed goal position (5 blocks, 15 starting and 5 release positions), with positions obtained from a previous calibration and marked on the table via printed outlines (Fig. 1a). The study was conducted with nine participants, each recording ten sets.

3 Modalities

We implement three video-based and one marker-based motion capture techniques, with temporal synchronisation and a common coordinate frame. We

adopt the *COCO-WholeBody* layout [16] (133 landmarks). Targeting a six-degree-of-freedom robot arm, we track a single arm following Gutzeit et al. [9] using four keypoints (back, shoulder, elbow, wrist) plus two on the thumb and index finger to capture the grasp, similar to Ren et al. [21]. A virtual end-effector point is added between the index finger and thumb (Figure 1c) as a potential robot target in the embodiment mapping step. Morphological differences between the marker setup and the keypoint locations (Figure 1b) are taken into account in the evaluation.

3.1 Marker-Based Motion Capture

The marker-based motion capture uses the *Qualisys* product suite, which provides automatic human pose estimation from passive markers and serves as the ground truth due to its high accuracy. The video-based modalities are translated into the *Qualisys* coordinate frame via a static frame constructed from multiple markers attached to the depth camera.

3.2 Video-Based Motion Capture

All three video-based techniques use data captured via a *ZED stereo camera*.

Direct 3D Pose Estimation: End-to-end models map an image to 3D coordinates, avoiding erroneous intermediate steps but remaining susceptible to background and environmental variations, and the vast solution space makes suboptimal solutions likely [20]. We use RTMW [14] with the RTMDet bounding-box detector [18], processing each frame independently.

Depth-Camera-Based 3D Pose Estimation: Mapping 2D keypoints to 3D via the per-pixel depth value is affected by occlusions, since the estimate is placed on the occluding object rather than on the body [5, 6]. Even without occlusion, a small offset occurs as the estimate sits on the body surface. We implement the naive 2D-to-3D projection using the 2D version of RTMW [14] with RTMDet [18]. To our knowledge no pose-refinement model with the *COCO-WholeBody* layout exists, and the naive approach performs adequately for this study.

Pose-Lifter-Based two-stage Pose Estimation: Pose-lifting lifts independent 2D keypoints to 3D without an external depth reference, extrapolating the depth axis from a single view. The 2-stage pipeline [12] builds on accurate 2D keypoint estimators. We adopt the *PAFUSE* model [22], which extends pose-lifting to the *COCO-WholeBody* dataset.

4 Data Collection

Data Synchronisation: The marker- and video-based pipelines are synchronized via nearest-neighbor down-sampling of the high-frequency marker timestamps to the low-frequency video timestamps, resulting in a negligible synchronisation error of $\Delta t_{sync} \approx 2.6 \pm 0.15$ ms.

Coordinate Transform: Pose estimation models output either person-centric or camera-centric coordinates. Most models without a depth reference are person-centric, but imitation learning requires a common coordinate system for pose and object. We bridge this gap with a Perspective-n-Point (PnP) projection from the estimated 2D image keypoints to the 3D object keypoints, yielding the camera pose relative to the person-centric output. PnP only needs RGB and a calibrated camera intrinsic K but is sensitive to calibration errors and 2D–3D mismatches. We therefore also implement a second approach that aligns the model output with the depth-map output via procrustes analysis [1, 8], combining the advantages of both modalities since the model output handles occlusions better than the depth map.

5 Evaluation Metrics

The primary contribution of this work is a set of human pose estimation data quality metrics decoupled from any single imitation learning algorithm. Traditional metrics measure whether an agent has learned the intended behaviour [13, 19], but they confound the algorithm and the record- and embodiment-mapping steps [3]. We instead evaluate the data on qualities that are generally important for imitation learning without committing to a particular technique.

Pose Estimation Quality: The most common metric, the *Mean Per Joint Position Error* (*MPJPE*), cannot be used here due to the morphology differences, but we include its derivatives *MPJVE* and *MPJAE* (velocity and acceleration). Since embodiment mapping (e.g., inverse kinematics) reduces the impact of keypoints other than root and end-effector, we additionally consider the velocity error of the virtual end-effector point and the coefficient of determination R^2_{speed} for the velocity difference [17]:

$$R^2_{speed} = 1 - \frac{\sum_i \|v_{marker}(i) - v_{video}(i)\|^2}{\sum_i \|v_{marker}(i) - \bar{v}_{marker}(i)\|^2} \tag{1}$$

where v_{marker} and v_{video} are the combined velocities in the x- and y-direction. Bone length deviation ΔL measures temporal consistency, as it should not vary across time. Due to the morphology differences we report the per-technique standard deviation and test whether the video-based deviation is significantly larger than the ground truth via a one-sided Wilcoxon signed rank test [25].

Trajectory Comparison: End-effector trajectories incorporate both path and velocity information. The static offset due to the morphology difference is removed via an initial spatial transform aligning the modalities at the first frame of each video [24]. The distance is measured both as the root mean square distance and as a normalized dynamic time warping [24], capturing the spatio-temporal and spatial relationships, respectively. We additionally compute the distance correlation *dCor* [23] and the trajectory accuracy $R^2_{trajectory}$ [17]:

$$R^2_{trajectory} = 1 - \frac{\sum_i \|p_m(i) - p_v(i)\|^2}{\sum_i \|p_m(i) - \bar{p}_m(i)\|^2} \tag{2}$$

where p_m and p_v are the marker- and video-based end-effector positions. $R^2_{trajectory}$ captures both path and velocity differences, while $dCor$, due to its double normalization [23], captures only the path correlation.

Behaviour Separability: This metric quantifies how well an agent could distinguish different actions a_n [13]. Behaviour distinction is captured by the end-effector position at the moment of grasping a block. We fit a K-Means clustering on the per-participant 2D end-effector positions projected onto the table plane at the grasping frame, and compare the predicted cluster labels against the true block labels, with a confidence interval from the bootstrapping approach of Champagne et al. [4].

Grasp Characterization: We use three metrics. The first is the 2D-plane distance from the end-effector to the closest point on the relevant shape outline at the grasping frame, tested against the marker-based modality via a two-sided Wilcoxon signed rank test [25]. The second is the target position error $RMSE_{pos}$ [17], the $RMSE$ of the L^2-norm between marker- and video-based end-effector positions at the grasping frame. The third is the approach-vector angle error $\Delta\alpha$, where the approach vector is orthogonal to the index-thumb vector and directed from the wrist toward the end-effector.

6 Results and Discussion

Full results are in Table 1.

Pose Estimation Quality: The depth-based approach captures the ground truth best and is the only modality whose velocity and acceleration metrics would make it usable for imitation learning. The pose lifter has the smallest bone length deviation ΔL, likely due to soft-tissue artifacts on the marker side [10, 2]. PnP projection markedly degrades the speed accuracy of the direct and pose lifter estimations, as it assumes a 2D-3D correspondence that does not

Table 1: Comprehensive experimental results for all modalities. Metrics marked n.a. are not applicable to the specific evaluation, while / indicates values exceeding the display range.

	Pose Estimation Quality						**Behaviour Separability**	
	Pose		End-Effector		Bone Var. (cm)			
Modality	MPJVE ($\frac{m}{s}$)	MPJAE ($\frac{m}{s^2}$)	Vel. Acc. ($\frac{m}{s}$)	Speed Acc.	ΔL (cm)	p-value	Mean Acc.	$CI_{95\%}$
Marker-based	n.a.	n.a.	n.a.	n.a.	0.28	n.a.	0.98	[0.89, 1.00]
Depth mapping	0.143 ± 0.049	1.089 ± 0.424	0.081 ± 0.045	0.75	0.45	0.027	0.94	[0.85, 1.00]
Pose Lifter (raw)	0.290 ± 0.048	1.472 ± 0.275	0.336 ± 0.143	0.26	0.19	0.844	0.67	[0.53, 0.82]
+ PnP projection	0.345 ± 0.119	2.556 ± 0.941	0.296 ± 0.232	−4.45	n.a.	n.a.	0.69	[0.56, 0.82]
+ procrustes analysis	0.164 ± 0.046	1.244 ± 0.407	0.116 ± 0.068	0.30	n.a.	n.a.	0.94	[0.82, 1.00]
Direct Est. (raw)	0.516 ± 0.274	3.924 ± 2.370	0.444 ± 0.385	−21.12	10.04	0.008	0.55	[0.44, 0.67]
+ PnP projection	/	/	2.261 ± 1.575	−554.69	n.a.	n.a.	0.57	[0.45, 0.68]
+ Procrustes analysis	/	/	0.552 ± 0.388	−20.67	n.a.	n.a.	0.71	[0.56, 0.89]

	Trajectory Dist.		**Trajectory Similarity**		**Shape Proximity**		**Grasp Metric**	
Modality	$RMSE_{lsed}$ (cm)	dtw_{norm} (cm)	R^2	$dCor$	Mean (m)	p-value	Distance (m)	Angle
Marker-based	n.a.	n.a.	n.a.	n.a.	0.007	n.a.	-	-
Depth mapping	4.26	1.06	0.879	0.933	0.032	0.0039	(0.064 ± 0.023)	35.61°
Pose Lifter (projection)	17.89	2.23	-1.165	0.956	0.367	0.0039	(0.442 ± 0.194)	34.76°
Pose Lifter (procrustes)	6.10	1.28	0.766	0.988	0.031	0.0039	(0.069 ± 0.023)	36.30°
Direct Est. (projection)	87.21	4.89	-56.653	0.752	0.661	0.0039	(0.754 ± 0.508)	53.49°
Direct Est. (procrustes)	13.08	1.86	-0.275	0.94	0.075	0.0039	(0.120 ± 0.067)	53.63°

hold here. Procrustes alignment improves the result mostly by averaging with the depth-map error.

Trajectory Distance: The $RMSE_{lsed}$ of the depth mapping is comparable to state-of-the-art MPJPE $\approx$ 5cm [14, 22, 11], while the two state-of-the-art models suffer from the coordinate system translation. $R^2_{trajectory}$ is adequate only for depth mapping. *dCor* is satisfactory for all modalities except the projected direct estimation, and notably exceeds depth mapping for the pose lifter, indicating that the path is well described once offset and velocity inconsistencies are removed. An imitation learning agent could thus learn a close representation of the trajectories from depth mapping and aligned pose lifter outputs.

Behaviour Separability: Figure 2 shows that the marker-based and depth mapping approaches both score near the maximum, demonstrating that the end-effector estimation is accurate enough to distinguish actions a. The pose lifter and direct estimation reach satisfactory accuracy only when aligned with the depth map via procrustes analysis. Their raw and projected variants remain too noisy.

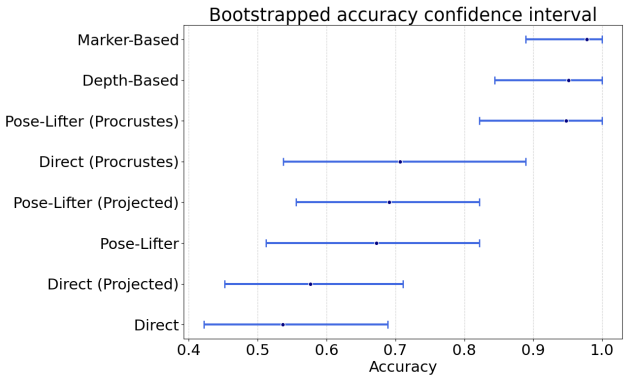


Fig. 2: Mean accuracy and 95% CI of bootstrapped K-Means clustering.

Grasp Characterization: The marker-based approach reaches a mean end-effector to shape distance of about 7 mm, against 3.2 cm for the depth mapping, in line with state-of-the-art MPJPE of $\approx$ 5 cm. The best end-effector position error is $\approx$ 6.5 cm with an angle error of $\sim 36°$. Such an end-effector distance would likely prevent an imitation learning agent from properly grasping the objects.

Conclusion: Depth-based estimation is a viable alternative to marker-based capture for trajectory-heavy imitation, but the evaluated video-based modalities cannot yet represent fine-grained tasks. The limiting factor is not pose estimation accuracy alone but also coordinate system and temporal consistency. Future work should explore depth-aware pose lifters or hybrid imitation and reinforcement learning, and extend the framework beyond this controlled-lab scenario.

Acknowledgments. This research was partially financed by the German Federal Ministry of Research, Technology and Space (BMFTR) through the ActGPT project, grant number 01IW25002, and partially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

Disclosure of Interests. The authors have no competing interests.

References

[1] Devrim Akca. *Generalized Procrustes Analysis and Its Applications in Photogrammetry*. June 1, 2003.

[2] Andrea Ancillao, Erwin Aertbeliën, and Joris De Schutter. "Effect of the Soft Tissue Artifact on Marker Measurements and on the Calculation of the Helical Axis of the Knee during a Squat Movement: A Study on the CAMS-Knee Dataset." In: *Medical Engineering & Physics* 110 (Dec. 1, 2022), p. 103915. ISSN: 1350-4533. DOI: 10.1016/j.medengphy.2022.103915. URL: https://www.sciencedirect.com/science/article/pii/S1350453322001631 (visited on 09/11/2025).

[3] Brenna D. Argall et al. "A Survey of Robot Learning from Demonstration". In: *Robotics and Autonomous Systems* 57.5 (May 31, 2009), pp. 469–483. ISSN: 0921-8890. DOI: 10.1016/j.robot.2008.10.024. URL: https://www.sciencedirect.com/science/article/pii/S0921889008001772 (visited on 05/15/2025).

[4] Catherine Champagne et al. "A Bootstrap Method for Assessing Classification Accuracy and Confidence for Agricultural Land Use Mapping in Canada". In: *International Journal of Applied Earth Observation and Geoinformation* 29 (June 1, 2014), pp. 44–52. ISSN: 1569-8432. DOI: 10.1016/j.jag.2013.12.016. URL: https://www.sciencedirect.com/science/article/pii/S0303243414000026 (visited on 09/17/2025).

[5] Andrea D'Eusanio et al. "RefiNet: 3D Human Pose Refinement with Depth Maps". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2020 25th International Conference on Pattern Recognition (ICPR). Jan. 2021, pp. 2320–2327. DOI: 10.1109/ICPR48806.2021.9412451. URL: https://ieeexplore.ieee.org/document/9412451 (visited on 09/18/2025).

[6] Mallika Garg, Debashis Ghosh, and Pyari Mohan Pradhan. "OccRobNet: Occlusion Robust Network for Accurate 3D Interacting Hand-Object Pose Estimation". In: *2025 National Conference on Communications (NCC)*. 2025, pp. 1–6. DOI: 10.1109/NCC63735.2025.10983334.

[7] Nathan Gavenski et al. *A Survey of Imitation Learning Methods, Environments and Metrics*. July 30, 2024. DOI: 10.48550/arXiv.2404.19456. arXiv: 2404.19456 [cs]. URL: http://arxiv.org/abs/2404.19456 (visited on 03/11/2025). Pre-published.

[8] J. C. Gower. "Generalized Procrustes Analysis". In: *Psychometrika* 40.1 (Mar. 1, 1975), pp. 33–51. ISSN: 1860-0980. DOI: 10.1007/BF02291478. URL: https://doi.org/10.1007/BF02291478 (visited on 09/16/2025).

[9] Lisa Gutzeit et al. "The BesMan Learning Platform for Automated Robot Skill Learning". In: *Frontiers in Robotics and AI* 5 (May 31, 2018), p. 43. ISSN: 2296-9144. DOI: 10.3389/frobt.2018.00043. URL: https://www.frontiersin.org/article/10.3389/frobt.2018.00043/full (visited on 05/15/2025).

[10] Lorna Herda et al. "Using Skeleton-Based Tracking to Increase the Reliability of Optical Motion Capture". In: *Human Movement Science* 20.3

(June 1, 2001), pp. 313–341. ISSN: 0167-9457. DOI: 10.1016/S0167-9457(01)00050-1. URL: https://www.sciencedirect.com/science/article/pii/S0167945701000501 (visited on 09/11/2025).

[11] Karl Holmquist and Bastian Wandt. "DiffPose: Multi-hypothesis Human Pose Estimation Using Diffusion Models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 15931–15941. DOI: 10.1109/ICCV51070.2023.01464.

[12] Chih-Hsiang Hsu and Jyh-Shing Roger Jang. "Enhancing 3D Human Pose Estimation with Bone Length Adjustment". In: *Computer Vision – ACCV 2024*. Springer, 2024, pp. 242–257. DOI: 10.1007/978-981-96-0885-0_14.

[13] Xiaogang Jia et al. "Towards Diverse Behaviors: A Benchmark for Imitation Learning with Human Demonstrations". In: (Oct. 13, 2023). URL: https://openreview.net/forum?id=6pPYRXKPpw (visited on 03/19/2025).

[14] Tao Jiang, Xinchen Xie, and Yining Li. *RTMW: Real-Time Multi-Person 2D and 3D Whole-body Pose Estimation*. July 11, 2024. DOI: 10.48550/arXiv.2407.08634. arXiv: 2407.08634 [cs]. URL: http://arxiv.org/abs/2407.08634 (visited on 03/11/2025). Pre-published.

[15] Xinkai Jiang et al. "A Comprehensive User Study on Augmented Reality-Based Data Collection Interfaces for Robot Learning". In: *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '24. New York, NY, USA: Association for Computing Machinery, Mar. 11, 2024, pp. 333–342. ISBN: 979-8-4007-0322-5. DOI: 10.1145/3610977.3634995. URL: https://dl.acm.org/doi/10.1145/3610977.3634995 (visited on 09/17/2025).

[16] Sheng Jin et al. "Whole-Body Human Pose Estimation in the Wild". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 196–214. ISBN: 978-3-030-58545-7. DOI: 10.1007/978-3-030-58545-7_12.

[17] A. Lemme et al. "Open-Source Benchmarking for Learned Reaching Motion Generation in Robotics". In: *Paladyn, Journal of Behavioral Robotics* 6.1 (Mar. 10, 2015). ISSN: 2081-4836. DOI: 10.1515/pjbr-2015-0002. URL: https://www.degruyterbrill.com/document/doi/10.1515/pjbr-2015-0002/html (visited on 09/24/2025).

[18] Chengqi Lyu et al. *RTMDet: An Empirical Study of Designing Real-Time Object Detectors*. Dec. 16, 2022. DOI: 10.48550/arXiv.2212.07784. arXiv: 2212.07784 [cs]. URL: http://arxiv.org/abs/2212.07784 (visited on 09/20/2025). Pre-published.

[19] Debasmita Mukherjee et al. "A Survey of Robot Learning Strategies for Human-Robot Collaboration in Industrial Settings". In: *Robotics and Computer-Integrated Manufacturing* 73 (Feb. 1, 2022), p. 102231. ISSN: 0736-5845. DOI: 10.1016/j.rcim.2021.102231. URL: https://www.sciencedirect.com/science/article/pii/S0736584521001137 (visited on 03/27/2025).

[20] Rama Neupane, Kan Li, and Tesfaye Boka. "A Survey on Deep 3D Human Pose Estimation". In: *Artificial Intelligence Review* 58 (Nov. 25, 2024). DOI: 10.1007/s10462-024-11019-3.

[21] Juntao Ren et al. "Motion Tracks: A Unified Representation for Human-Robot Transfer in Few-Shot Imitation Learning". In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. 2025, pp. 8802–8810. DOI: 10.1109/ICRA55743.2025.11128834.

[22] Nermin Samet et al. "PAFUSE: Part-Based Diffusion for 3D Whole-Body Pose Estimation". In: *Computer Vision – ECCV 2024 Workshops*. Springer, 2024, pp. 151–169. DOI: 10.1007/978-3-031-91575-8_10.

[23] Gábor J. Székely and Maria L. Rizzo. "Brownian Distance Covariance". In: *The Annals of Applied Statistics* 3.4 (Dec. 1, 2009). ISSN: 1932-6157. DOI: 10.1214/09-AOAS312. arXiv: 1010.0297 [stat]. URL: http://arxiv.org/abs/1010.0297 (visited on 09/25/2025).

[24] Yaguang Tao et al. "A Comparative Analysis of Trajectory Similarity Measures". In: *GIScience & Remote Sensing* 58.5 (July 4, 2021), pp. 643–669. ISSN: 1548-1603. DOI: 10.1080/15481603.2021.1908927. URL: https://www.tandfonline.com/doi/full/10.1080/15481603.2021.1908927 (visited on 09/09/2025).

[25] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6 (Dec. 1945), pp. 80–83. DOI: 10.2307/3001968.