Computerised Assessment of Motor Imitation As a Scalable Method for Distinguishing Children With Autism

Supplementary Information

This PDF file includes:

Supplementary Information Text

Participants

A thorough set of inclusion and exclusion criteria were applied for children in the ASC group. All clinical assessments, including those for autism diagnosis, were conducted by trained psychology associates, and the assessments were supervised and verified by a child neurologist (senior author S.H.M.). Participants were screened to exclude individuals with co-occurring neurological or medical conditions that might confound the results including (i) known genetic disorder (e.g., NF1, tuberous sclerosis), acquired neurologic disease (e.g., stroke, tumour), cerebral palsy, history of severe head injury, intracranial pathology or significant dysmorphology, (ii) history of seizures or confirmed diagnosis of epilepsy, (iii) any progressive (e.g., neurodegenerative) neurological disorder, (iv) history of head injury resulting in prolonged loss of consciousness, (v) active psychosis, major depression, bipolar disorder, conduct disorder, or adjustment disorder. Presence and history of psychiatric diagnoses were assessed using a comprehensive standardised parent interview, the Kiddie Schedule for Affective Disorders and Schizophrenia for School-Aged Children – Lifetime Version (KSADS). Children with co-occurring anxiety and attention deficit hyperactivity disorder (ADHD) were included due to high rates of co-occurrence of these disorders in autism. Children on stimulant medication had the medication held on the day before and on the day of testing.

Children were included in the TD group if they: (1) did not meet published cut-off criteria for autism spectrum disorders on the Social Responsiveness Scale (SRS-2), (2) did not have a history of ADHD, developmental disorder, or a psychiatric disorder based on maternal and child responses from the KSADS, and (3) did not have any immediate family members (sibling, parent) with ASC.

Procedure

The study took place in a room equipped with two Kinect Xbox cameras and a large TV screen (see Figure S1 for a schematic of the room and the video stimuli). The data presented in this paper was collected as part of a larger study examining motor imitation skills in ASC. As part of this larger study, some of the children were recorded using the Vicon Motion Capture Systems in addition to Kinect Xbox. As such, those children, such as the one shown in Figure S1, wore some sensors strapped around their arms, legs, waist and hips. The Vicon data were not analysed for the current paper. The Kinect data reported in this paper did not require children to wear any special clothing or sensors.

The entire task, only a part of which is reported in this paper, comprised of 14 movement trials that used three different movement sequences. One of these sequences was presented only twice (sequence-1): once at the beginning and once at the end of the session (Trials 1a and 1b in the paper). The other two sequences were presented at varying speeds across 12 trials, starting with 100% and then gradually slowing down before finishing at 100% speed again. To avoid potential confound of different speeds, in the paper, we reported only the very first trial that the children performed at 100% speed (sequence-2, Trial 2a in the paper) and its post-training repetition at 100% speed (sequence-2, Trial 2b in the paper). The whole procedure took around 40 minutes including set-up, task instructions and breaks to avoid fatigue.

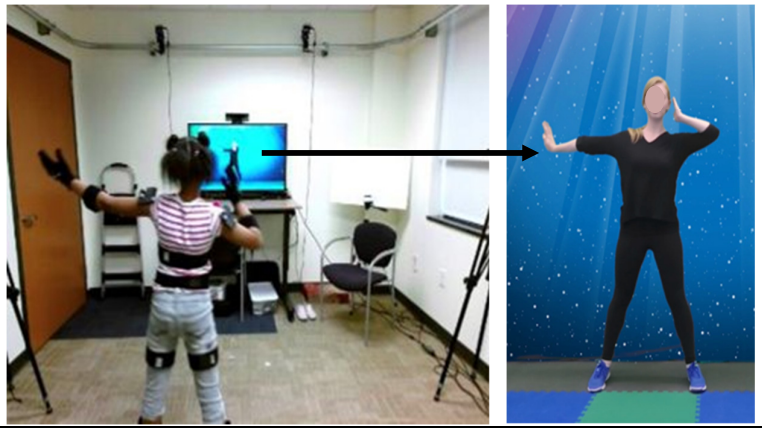


Figure S1. Schematic of the study room (left) and a screen shot of the stimuli showing the model performing a move (right).

Data Processing and Extraction

The motion tracking data collected through the Kinect Xbox depth cameras was subjected to several processing steps before the x-y-z coordinates of the joints for each time frame were extracted.

During testing, each trial's data were saved as two separate motion files, one from the front camera and one from the rear camera. These two files were merged using iPi Recorder versions 3 and 4 with a calibration file that was recorded on day of testing. The merged files were then processed using iPi MoCap Studio software versions 3 and 4. This processing entailed imposing a skeleton onto the depth data of the child obtained through the Kinect Xbox cameras throughout the whole movement sequence using the "Track" feature of the software. The skeleton was set to be the same height as the child, as measured on day of testing. However, due to problems with this automated tracking (i.e., the skeleton still not aligning with the depth data in the tracked version), all videos were reviewed by at least two coders who used a combination of manual editing and the automated tracking feature of the software.

The following settings were enabled on the iPi Motion Capture software before tracking: Head tracking, Foot tracking, Ground collisions, Shoulder: calculated from arm position, Spine: stiff lower spine, Tracking resolution: high, Trajectory filter: 1.

After the depth data was ensured to align with the skeleton for the entirety of the movement sequence, the motion data was exported using the Biomech add-on of the iPi MoCap Studio software. The x-y-z coordinates of data from the following 20 joints were exported: Hip, LowerSpine, MiddleSpine, Chest, Neck, Head, RClavicle, RShoulder, RForearm, RHand, LClavicle, LShoulder, LForearm, LHand, RThigh, RShin, RFoot, LThigh, LShin and LFoot. Considering issues with unreliability of the software in tracking the extremities and body parts intersecting with the ground, we excluded the joints LToe, EffectorLToe, RToe and EffectorRToe. Further, due to the negligible informativeness of the eye or the effector head, we excluded the joints LEye, REye and EffectorHead. The coordinate system was set to: "Relative to center of mass" and the coordinates were taken in "Centimeter" units.

Human Observation Coding (HOC) Scheme

Each movement sequence was split into distinct movement types. Sequence-1 had 14 movement types, and sequence-2 had 18 movement types. Each movement type was defined as a coherent movement unit, which could involve simultaneous movement

of different limbs and repetitions of a pattern. See Figure S2 for an example of a movement type. In sequence-1, the number of elements that made up a movement type, which can also be considered as that movement type's length, varied between 5 and 29, with an average of 12.00 (SD = 7.48) elements. For sequence-2, the number of elements ranged between 2 and 19, with an average of 12.57 (SD = 5.45) elements.

	Movement Type #1	Complete	Reverse Side
1	Right arm down on the side	1	0
2	Step left leg out to left	1	-0.5
3	Sweep left arm away from the body to the left	0	0
4	Keep left arm in front of body while sweeping	0	0
5	Step right leg in, feet together	1	0
6	Left arm down on the side	1	0
7	Step right leg out to left	0	0
8	Sweep right arm away from the body to the left	1	0
9	Keep right arm in front of body while sweeping	0	0
10	Step left leg in, feet together	1	-0.5
11	**Repetition** – repeated more than twice? [REVERSE ITEM]	-1	N/A
	Total Score for Movement Type #1	**4 (out of 10)**	

Figure S2. Snippet from the HOC scheme, depicting one movement type from sequence-1. The scoring is for illustrative purposes only. In this example, the total score obtained from this movement type would be equal to 4, which would then be normalised by the maximum possible score, yielding a score of 0.4.

Each element of a movement type was given a score of 1 or 0 depending on whether it was completed or not, respectively. In addition, for each element that was performed on the reverse side of the body (e.g., using right leg instead of left leg), a score of 0.5 was deducted. Finally, some movement types were comprised of repetitions of an action pattern. For instance, the example in Figure S5 has two repetitions of an action pattern because elements 1-5 are repeated on the reverse side of the body in elements 6-10. If the children performed more repetitions than what was demonstrated, a score of 1 was deducted; this deduction was made only once per movement type.

Children's total HOC score was a combination of their score of the positive items (s_{pos}) and the negative items (s_{neg}). s_{pos} was comprised of scores given to completed elements, which could be either a 1 (completed) or 0 (not completed). s_{neg} was comprised of scores given to incorrectly performed elements (i.e., "reverse side" and "repetition" items), which took values of 0 (not performed on the reverse side or repeated more than demonstrated), -0.5 (performed, but on the reverse side) or -1 (performed, but repeated for more times than demonstrated). A child's HOC score for a given movement type was the summation of s_{pos} and s_{neg}, divided by the maximum possible score a child could obtain from that movement type. The normalisation was done due to the large variance in the length of movement types. Total HOC scores were equal to the average of scores calculated per movement type (M). If the movement types had k_m elements each, where ($m = \{1,2,\dots,M\}$), scm corresponded to child c's score in movement type m. A child c's total HOC score (HOC_c) was calculated by averaging this child's normalised movement type scores. Note that repetition items were not considered as elements of a movement type and hence were not counted towards the variable k_m. The equation used to calculate the children's overall HOC score was as follows:

$$HOC_c = \frac{1}{M}\sum_{m=1}^{M} s_{cm} = \frac{1}{M}\sum_{m=1}^{M} \frac{s_{pos}{}_{cm} + s_{neg}{}_{cm}}{k_m} \quad \text{(eq.1)}$$

Computerised Assessment of Motor Imitation (CAMI) Algorithm

Problem formulation

Let $X_A \in \Re^{K \times T_A}$ be the matrix that contains 3D positions of J joints ($K = 3J$) of a model performing a dance-like sequence for T_A time steps. The sequence performed by the model is composed of M consecutive movement types, such that movement type m is performed in a known interval $\Omega_m \subseteq [1,2,\ldots,T_A]$. The subsets $\Omega_m$with $m = \{1,\ldots,M\}$ form a partition of the interval $\Omega = [1,2,\ldots,T_A]$.

Let $X_c \in \Re^{K \times T_c}$ be the matrix that contains 3D positions of J joints ($K = 3J$) of child c imitating the movements of the model for T_c time steps. The dataset $\mathcal{D} = \{X_c\}_{c=1}^{N}$ contains imitation instances from N children.

We considered the problem of assigning a score $0 \leq s_c \leq 1$ to each $X_c \in \mathcal{D}$, such that it reflects the quality of this child's imitation. In particular, we aim to find a function $f(X_A, X_C) = s_c$ that accurately and automatically assesses how well child c imitated the model.

To achieve this goal, we designed a process that was divided in three steps: (1) pre-processing, (2) feature extraction, and (3) score computation. In the following, we provide a step-by-step guideline to enable the replication of our method which was implemented using Matlab softare version 2018a.

1. Pre-processing

As emphasised in previous works [1] [2], pre-processing of skeleton data is a fundamental step to perform any type of body motion analysis. When comparing two sequences, pre-processing must account for at least (i) differences in the position of the subjects in space, (ii) differences in body size, and (iii) differences in the spatial orientation of the subjects.

To account for differences in the position of the subjects in space, we consider each participant's point of reference to be their own hip for every time step. To this end, the position of the hip was subtracted from the position of every other joint at every time step.

Since the body size and limb length of adults and children is expected to be different, it is important to map the model's and the children's data to a skeleton of a fixed size. The skeleton diagram used for this is shown in Figure S3; the dimensions of the depicted skeleton were defined by the average length of the model's body segments across time steps. Given this skeleton as a reference and using the hip joint as the root, we applied the skeleton normalisation algorithm proposed in [3].

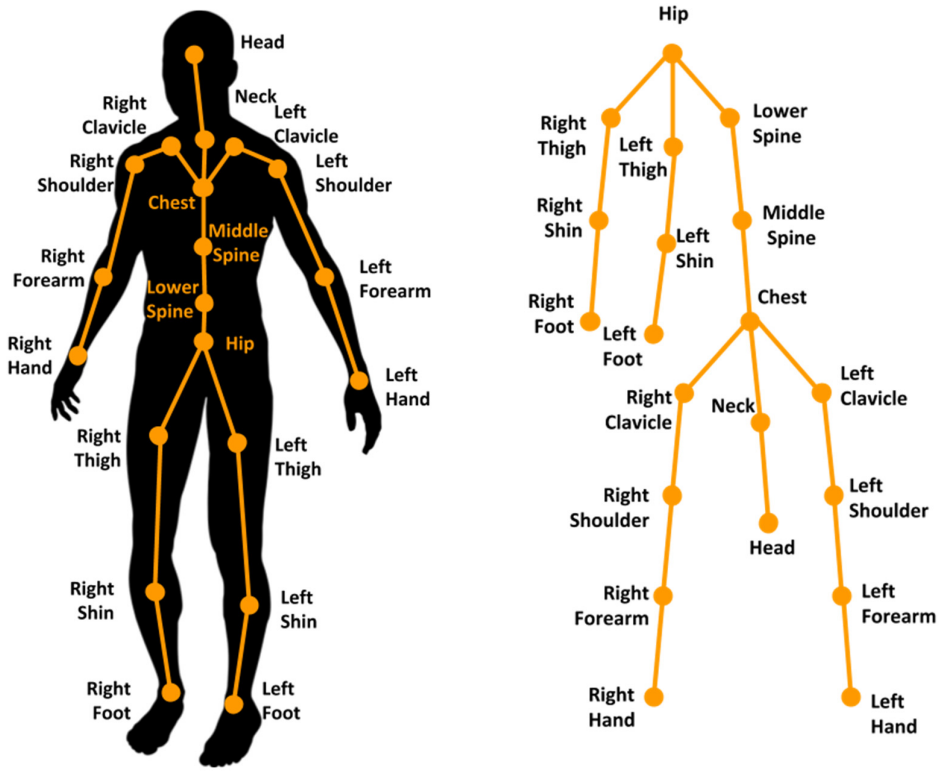


Figure S3. Representation of the skeleton. Diagram of the localisation and connections of the 20 joints used (left) and a tree-like representation of the same joints used to perform breadth-first search for the skeleton normalisation algorithm (right).

Next, we corrected for differences in the initial spatial orientation of the child. A difference in the initial orientation of the child with respect to the initial orientation of the model could stem from the fact that the children were allowed to move in a wide, unrestricted space. Hence, if uncorrected, even a small difference in initial orientation could unduly affect the children's imitation performance measure in subsequent time frames. To prevent this type of artefact, we aligned the orientation of the child and the model based on information from their initial position. Given that the the main variation between the model's and the child's poses was expected to be a rotation around the vertical axis, the vector between one's left and right shoulders, which is orthogonal to the vertical axis, is considered to be representative of spatial orientation. This orientation correction was made only based on the first frame, because changes in children's spatial orientation during the imitation task were considered as a sign of poor imitation performance.

Let v_c be the vector that describes the difference between left and right shoulders' positions of child c in the first frame, and v_A be the equivalent for the model. The angle θ_c between both vectors was obtained by

$$\theta_c \equiv cos^{-1}\left(\frac{<v_c, v_A>}{||v_c||_2 ||v_A||_2}\right), \qquad \text{(eq.2)}$$

where $<\cdot,\cdot>$ denotes dot product and $||\cdot||_2$ corresponds to the Euclidean norm of vectors. The rotation matrix $R(\theta_c)$ is defined by

$$R(\theta_c) = \begin{bmatrix} \cos(\theta_c) & 0 & -\sin(\theta_c) \\ 0 & 1 & 0 \\ \sin(\theta_c) & 0 & \cos(\theta_c) \end{bmatrix}. \quad \text{(eq.3)}$$

Let $x_c^j(t)$ be the vector that represents the hip-referenced and size-normalised position of joint j of child c in time step t. Then, for each time step and joint, the new coordinates are obtained by

$$\hat{x}_c^j(t) = x_c^j(t) \cdot R(\theta_c), \quad \text{(eq.4)}$$

where · denotes matrix multiplication. All pre-processed joints' positions of a child for a given time step, $\hat{x}_c^j(t) \in \Re^3$ for $j = 1,2,\ldots,J$, are concatenated to form a single vector $\hat{x}_c(t) \in \Re^{3J} = \Re^K$. Finally, the column vectors $\hat{x}_c(t)$ for $t = 1,2,\ldots,T_c$ are stacked together to obtain the matrix $X_c \in \Re^{K \times T_c}$ that contains the pre-processed joints' positions of child c. The same process is followed in order to obtain the pre-processed joints' positions of the model $X_A \in \Re^{K \times T_A}$.

2. Feature extraction

Unlike other comparative tasks, imitation must consider not only the shape of the movements performed, but also their timing relative to the model's movements [4] [5]. We aimed to capture both shape and timing differences by extracting three features based on dynamic time warping (DTW): one related to the spatial distance between the child's positions and the model's positions, and two others related to the timing differences between the child's and the model's sequence of movements.

Firstly, we compute DTW to find the optimal alignment path $\varphi^*(X_c, X_A)$ that minimises the Euclidean distance between the child's movements X_c and the model's movements X_A (details on the constraints and solvers of this optimisation problem can be found in [6]):

$$\varphi^*(X_c, X_A) = \underset{\varphi(X_c,X_A)}{\operatorname{argmin}} \sum_{(n_c,n_A)\in\varphi(X_c,X_A)} || \hat{x}_c(n_c) - \hat{x}_A(n_A)||_2. \quad \text{(eq.5)}$$

The distance minimised above considers each joint and each time step as equally important. This is problematic for our purposes of assessing imitation in a sequence of naturalistic movements because (1) some joints are more relevant than others for a given movement type, and (2) each movement type should contribute equally to the overall score regardless of the length of that particular movement type. To achieve this, we define the distance between X_c and X_A as:

$$d(X_c, X_A) = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{|\varphi_m|} \sum_{(n_c,n_A)\in\varphi_m} \sum_{j=1}^{J} \frac{{\alpha_j}^m}{\sqrt{3}} \left\| \hat{x}_c^j(n_c) - \hat{x}_A^j(n_A) \right\|_2 , \quad \text{(eq.6)}$$

where $\varphi_m \equiv \left\{ (n_c, n_A) \in \varphi^*(X_c, X_A)\ s.t.\ n_A \in \Omega_m \right\}$ corresponds to the subset of the optimal alignment path that corresponds to movement type m; Ω_m is the set containing the time steps in which the model performs movement type m; $\varphi^*(X_c, X_A)$ is the optimal temporal alignment path as defined in (eq. 5); $|\varphi_m|$ is the cardinality of the set; ${\alpha_j}^m$ is a scalar that represents the relevance of joint j in movement type m; $\hat{x}_c^j(n_c) \in \Re^3$ is the position of joint j of child c in time step n_c; $\hat{x}_A^j(n_A) \in \Re^3$ is the position of joint j of the model in time step n_A; J is the total number of joints; and M is the total number of movement types in the sequence.

To automatically assess the relevance of each joint for a given movement type, we follow the idea presented in [7], in which the relevance of each joint is characterised by its displacement during a given movement type. Then, we define the normalised displacement of joint j in movement type m as

$$D_j{}^m = \frac{\sum_{n_A \in \Omega_m} \left\| \hat{x}_A^j(n_A) - \hat{x}_A^j(n_A - 1) \right\|_2}{max_{\ j \in [1\ ,J\ ]} \left\{ \sum_{n_A \in \Omega_m} \left\| \hat{x}_A^j(n_A) - \hat{x}_A^j(n_A - 1) \right\|_2 \right\}}, \quad \text{(eq.7)}$$

from which the relevance factors $\alpha_j{}^m$ are computed using a sigmoidal transformation

$$\alpha_j{}^m = \frac{1 - e^{-\frac{D_j^m}{\sigma_D}}}{\sum_{j=1}^{J} \left(1 - e^{-\frac{D_j^m}{\sigma_D}} \right)}, \quad \text{(eq.8)}$$

where $\sigma_D$corresponds to the standard deviation of all normalised displacements, considering all joints and all movement types.

The distance $d(X_c, X_A)$ has an inverse relationship with imitation performance: the smaller the distance, the better the imitation performance. To make the interpretation of this variable more intuitive, the distances are transformed into scores by an exponential function

$$s_{dist}(X_c, X_A) = e^{-\lambda \frac{d^{\,2}(X_c, X_A)}{\sigma_d{}^2}}, \quad \text{(eq.9)}$$

where $\sigma_d{}^2$ is the variance of the distances, and λ is a parameter to be determined (see *Parameter Learning* section below for more details). After this transformation, the resulting s_{dist} score has a positive relationship with imitation performance, where higher scores indicate better imitation performance.

The optimal alignment path $\varphi^*(X_c, X_A)$ computed in (eq. 5) contains relevant information regarding the timing differences between the movements of child c and the model. Using the optimal alignment path $\varphi^*(X_c, X_A)$, we computed the percentage of time the child was delayed with respect to the model $t_{delay}(X_c, X_A)$ and the percentage of time the child moved in advance of the model $t_{adv}(X_c, X_A)$ by following the definitions presented in [8].

3. Score computation

The performance of child c is characterised by the score associated with spatial discrepancies $s_{dist}(X_c, X_A)$, the proportion of the time that the child was delayed with respect to the model $t_{delay}(X_c, X_A)$, and the proportion of time that the child moved in advance of the model $t_{adv}(X_c, X_A)$. The imitation performance coefficient $i(X_c, X_A)$ of child c is modelled as a weighted linear combination of these features, as shown in (eq. 10). The values of the weights $(W_{dist}, W_{delay}, W_{adv})$ were determined in a data-driven manner, as described in detail in the *Parameter Learning* section.

$$i(X_{c,}X_A) = W_{dist}\, s_{dist}(X_c, X_A) + W_{delay}\, t_{delay}(X_c, X_A) + W_{adv}\, t_{adv}(X_c, X_A). \quad \text{(eq.10)}$$

To ensure that the imitation scores of the children lie in a reasonable and interpretable range, we computed the imitation performance coefficient of a "best imitation" and a "worst imitation" scenario. The "best imitation" case (X_{best}) was

obtained by comparing the model's performance to the sequence performed by the lead experimenter (third author, R.R.), who had extensive experience observing the movements and analysing them through the creation of the HOC scheme and performing HOC. This "best imitation" scenario enabled establishing a reasonable upper bound to the performance of the children. The "worst imitation" case (X_{worst}) was defined as not performing any imitation, i.e., staying still for the whole duration of the sequence. This "worst imitation" scenario was quantified by repeating the model's position in the first time-step for the whole duration of the experiment and calculating the distance between this case and the model. All children's scores were then fit within the range of "worst imitation" and "best imitation" cases by

$$s_c = min\left\{max\left\{\frac{i(X_c, X_A) - i(X_{worst}, X_A)}{i(X_{best}, X_A) - i(X_{worst}, X_A)}, 0\right\}, 1\right\} \in [0,1]. \quad \text{(eq.11)}$$

Parameter Learning

The method described above requires finding a set of four parameters: $\lambda > 0$ to transform distances to scores in (eq.9), and $\{W_{dist}, W_{delay}, W_{adv}\}$ to define the performance coefficient in (eq.10). We calculated these parameters in two ways: (i) using 3-fold cross-validation, and (ii) using the whole dataset to learn a single set of parameters (more details can be found in the *Results* section below).

We assume there exist V datasets ($\mathcal{D}_v$ for $v = \{1,2,\dots,V\}$), with data from different sequences or imitation instances (e.g., first, second or third time that the children are exposed to a given sequence). These datasets have been coded by human observers; for every $X_c \in \mathcal{D}_v$, for $c \in \{1,2,\dots,N\}$ and $v = \{1,2,\dots,V\}$ there exists a corresponding HOC_c as given by (eq.1).

In order to find λ we do an exhaustive search in the range $[0, 0.1]$ and we choose the one that maximises the average correlation between distance scores s_{dist} and human observed codes HOC (average computed across datasets). Once λ has been set, we perform gradient ascent with a learning rate of $\epsilon = 0.01$ and random initialisation to find the 3-tuple $\{W_{dist}, W_{delay}, W_{adv}\}$ that maximises the average correlation (across datasets) between human observed codes HOC and the performance coefficients i as computed by (eq.10). The optimal parameters are normalised such that $W_{dist}^2 + W_{delay}^2 + W_{adv}^2 = 1$.

Results

1. *Learnt parameters of the CAMI method*

The model of the CAMI method has four parameters that need to be learnt from the data (i.e., λ and three weight parameters for *s_dist*, *t_delay and t_adv* as defined in the *Parameter Learning* section). To learn these parameters, we first applied 3-fold cross-validation by splitting the dataset into 3 groups and using these groups to form 3 folds, each one consisting of 2 groups for training and 1 group for testing. We then obtained the parameters from the training examples of each fold, yielding three sets of parameters. In addition, to learn a single set of parameters that can be used in future research, we used the whole data set. The parameters obtained in every instance of the cross-validation procedure and in the case where the whole dataset is used for training are reported in Table S1. The fact that for all cases $W_{dist} > 0$ confirms that children with higher distance scores get higher performance coefficients. Similarly, the fact that $W_{delay}, W_{adv} < 0$ is expected because timing differences with respect to the model are

indicative of poorer imitation performance, and hence lead to lower performance coefficients.

Table S1. Parameters learnt using 3-fold cross-validation and whole dataset analysis.

	λ	W_{dist}	W_{adv}	W_{delay}
Split				
1	0.0260	0.7825	-0.5051	-0.3641
2	0.0240	0.6542	-0.7203	-0.2306
3	0.0300	0.6601	-0.4583	-0.5952
Whole dataset	0.0270	0.7200	-0.4667	-0.5137

The CAMI scores obtained using 3-fold cross-validation and whole-data set were significantly similar to each other (**Trial 1a**: $r(43) = .99$, **Trial 1b**: $r(40) = .96$, **Trial 2a:** $r(46) =. 99$, **Trial 2b:** $r(39) =. 99$**,** all $ps < .0001$).

2. *Including time-related features improved CAMI algorithm's performance*

We checked whether including temporal information (t_adv and t_delay variables) in addition to the spatial distance information (s_dist) improved the CAMI model's correlation with the HOC scores. Figure S4 illustrates the findings for each trial and for the average of all three trials, which reveal higher (or similar, in case of Trial 2a) correlations. Notice that the distances alone (before maximising correlation with HOC), as well as the distances scores already yield good correlation (coefficients > .70) between the HOC and CAMI methods, while adding temporal information further improves the CAMI method's detection of imitation ability. This experiment supports the idea that temporal information is an important component when assessing imitation.

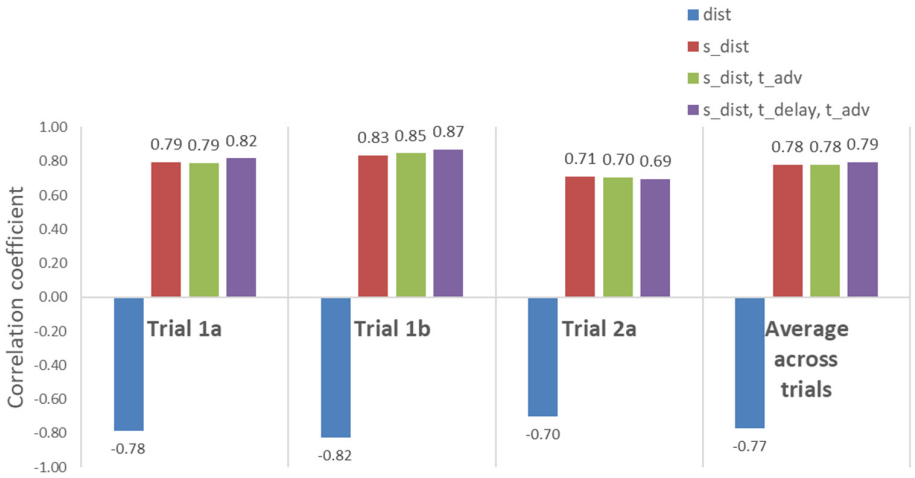


Figure S4. Correlation between HOC and CAMI scores in conditions when CAMI scores are calculated using only the distances (blue), s_dist variable (red), s_dist and t_adv variables combined (green), and s_dist, t_delay and t_adv variables combined (purple). Higher distances (blue) indicate worse imitation, while higher s_dist indicates better imitation. A correlation coefficient of 1 indicates perfect positive association, a correlation coeffieicnt of -1 indicates perfect negative association, while a correlation coefficient of 0 indicates no association.

3. *Continuous scores improved the CAMI model's diagnostic classification performance as compared to discrete scores*

There is a quantisation error associated with any discretisation process. Thus, in general, continuous variables contain more information than discrete ones. To illustrate this premise in the context of the imitation scores obtained from the HOC and CAMI methods, we discretised the scores and computed the area under the curve (AUC) of the receiver operating characteristic (ROC) curve as a measure of discriminative ability when classifying the subjects according to their diagnosis.

Figure S5 depicts the AUC of the ROC curve as we increase the number of discrete levels allowed in the discretisation process. In this curve, larger values indicate better discriminative ability. The discretisation is carried out in quantiles, which means that the scores are sorted in ascending order, and then the sorted continuous scores are divided into the desired number of discrete levels, each one containing approximately equal number of samples. As observed in Figure S5, although there is some variability inherent to empirical data, there is a clear trend in which discretising the scores into a smaller number of levels leads to poorer discriminative ability in terms of the AUC. This supports the idea that continuous scores are more expressive, and in this case are more informative than discrete scores to classify the participants into diagnostic groups.

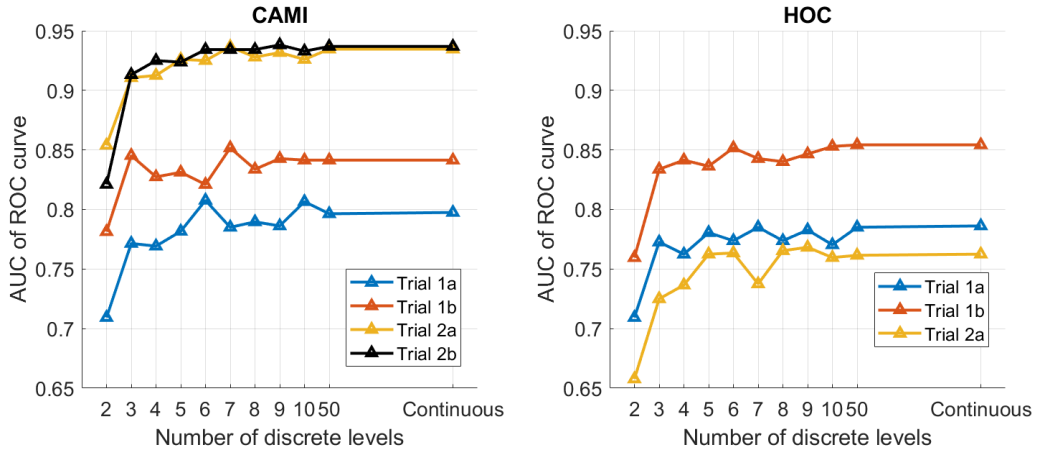


Figure S5. Area under the curve (AUC) of receiving operating characteristic (ROC) curves vs the number of discrete levels in which the continuous scores are discretised. The ROC curves are formed by true positive rates vs false positive rates as the classification threshold is varied. Larger AUC indicates better diagnostic ability of the method (best possible AUC is 1, meaning 100% true positives and zero false positives). "Continuous" indicates the AUC values when no discretisation is applied.

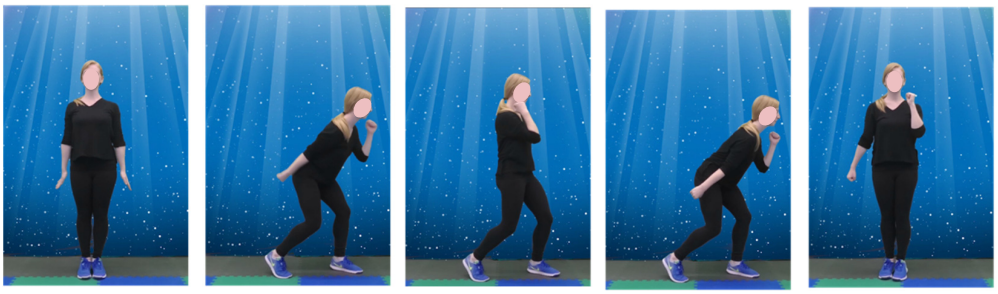
4. *CAMI method successfully detected important joints automatically*

In order to assess the validity of the CAMI method's automatic assessment of joint importance, we used the HOC scheme to manually determine the top five most important joints for each of the 18 movement types of sequence 1. This analysis was done blindly, without reference to the CAMI method. Then, the joint weights as determined by CAMI were ranked from highest (most important joint) to lowest (least important joint) for each movement type. The top five joints as determined by the HOC method and the CAMI method were compared. This comparison revealed that, on average, the two methods identified the same joints as among the most important five joints in 91.67% of the time (see Table S2).

Table S2. Percent agreement between HOC and CAMI in terms of the top five most important joints for each of the 18 movement types of sequence 1

Movement Type	Percent agreement	Movement Type	Percent agreement	Movement Type	Percent agreement
1	80%	7	80%	13	80%
2	100%	8	100%	14	100%
3	100%	9	80%	15	80%
4	100%	10	100%	16	100%
5	80%	11	100%	17	100%
6	100%	12	100%	18	100%
MEAN OVERALL AGREEMENT = 91.67%					

Figure S6 shows an example movement type and its top important joints as identified by the HOC and CAMI methods. When the two methods had disagreement, one of the top five joints as determined by the HOC method was not within the top five of the CAMI method. However, in all of those cases, that remaining joint was ranked among the top eight joints according to the CAMI method. Therefore, based on the results of this experiment, we can conclude that the automatic detection of important joints in the CAMI method is in line with human observations.



Top 5 joints – HOC: LHand, RHand, LForearm, RForearm, LFoot
Top 8 joints – CAMI: LHand, RHand, LForearm, RShoulder, LFoot, RForearm, LShin, Head

Figure S6. Snapshots of movement type #15 and the top joints as determined by the HOC and CAMI methods.

5. *Demographic and clinical characteristics of children correctly and incorrectly diagnosed by CAMI*

As reported in the main article, a Support Vector Machine (SVM) approach using the CAMI scores of children in three one-minute imitation trials could correctly diagnose the children into ASC vs TD groups with 87.2% accuracy. This meant that out of the 27 children with ASC, only two of them were misdiagnosed. In Table S3, we report how these two children differed from the correclt diagnosed children on a range of dimensions. As can be seen, the correctly diagnosed children were very similar to the incorrectly diagnosed children in terms of age, IQ and autism severity as assessed by SRS-2 and ADOS-2. Due to the large difference in the number of correctly (n = 25) and incorrectly diagnosed children (n = 2), we did not conduct a statistical comparison between the two groups.

Table S3. Means and standard deviations of demographic information and core autism symptoms for children correctly vs incorrectly diagnosed by SVM using CAMI scores.

	ASC – correctly diagnosed Mean (SD)	**ASC – not correctly diagnosed** Mean (SD)
Chronological age (years)	10.36(1.41)	11.90(1.24)
SRS-2 SCI sub-scale	73.21(7.50)	68.50(14.89)
SRS-2 RRB sub-scale	74.74(9.99)	77.50(6.36)
SRS-2 total score	74.84(7.53)	74.00(11.31)
ADOS-2 SA sub-scale	11.80(3.33)	10.00(--)*
ADOS-2 RRB sub-scale	3.75(2.02)	2.00(--)*
ADOS-2 total score	15.55(3.73)	12.00(--)*
WISC-V Full scale IQ	99.95(17.88)	97.00(1.41)
Gender (Boys/Girls)	18/2	2/0

* One of the two misdiagnosed children's ADOS data is unavailable, and hence the ADOS values reported here belong to a single child.

6. Replication of the results using 10-fold cross-validation scheme

To study the potential disadvantages of using a small number of folds in the k-fold cross-validation scheme, the model of the CAMI method was also trained in a 10-fold cross-validation scheme. The parameters obtained in every instance of the cross-validation procedure are reported in Table S4. The fact that for all cases $W_{dist} > 0$ confirms that children with higher distance scores get higher performance coefficients. Similarly, the fact that $W_{delay}, W_{adv} < 0$ is expected because timing differences with respect to the avatar are indicative of poorer imitation performance, and hence lead to lower performance coefficients.

Table S4. Parameters learnt using 10-fold cross-validation scheme.

	$\boldsymbol{\lambda}$	$\boldsymbol{W_{dist}}$	$\boldsymbol{W_{delay}}$	$\boldsymbol{W_{adv}}$
Split				
1	0.0300	0.7693	-0.4705	-0.4321
2	0.0310	0.7383	-0.4919	-0.4614
3	0.0250	0.6040	-0.1271	-0.7868
4	0.0220	0.7028	-0.5096	-0.4963
5	0.0270	0.6810	-0.5964	-0.4249
6	0.0270	0.7024	-0.5582	-0.4417
7	0.0250	0.6963	-0.5538	-0.4565
8	0.0280	0.7350	-0.4900	-0.4688
9	0.0300	0.7580	-0.4278	-0.4923
10	0.0300	0.7366	-0.5154	-0.4379

The CAMI scores obtained using 10-fold cross-validation and 3-fold cross-validation method were significantly similar to each other (**Trial 1a**: $r(43) = .98$, **Trial 1b**: $r(40) = .96$. **Trial 2:** $r(46) = .98$, all ps $< .0001$). As can be seen in Figure S7, the

correlations between CAMI and HOC methods in all three trials was significant, with a slight improvement being observed as compared to the 3-fold scheme.

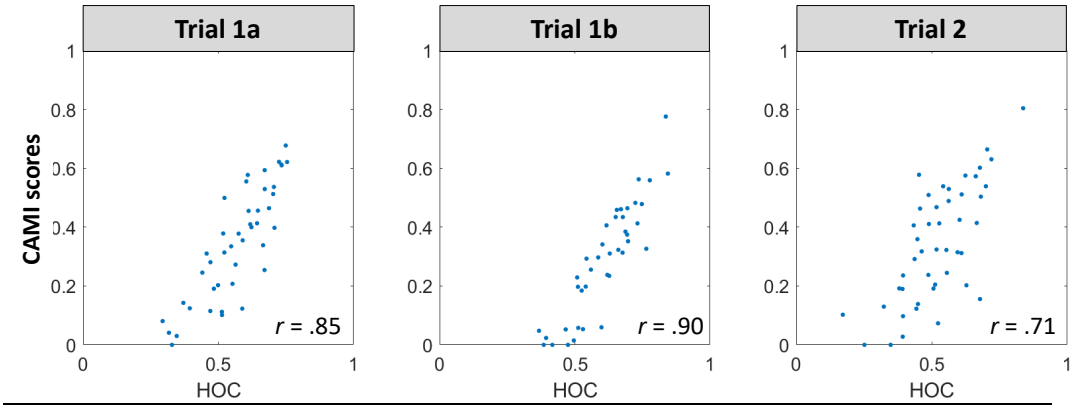
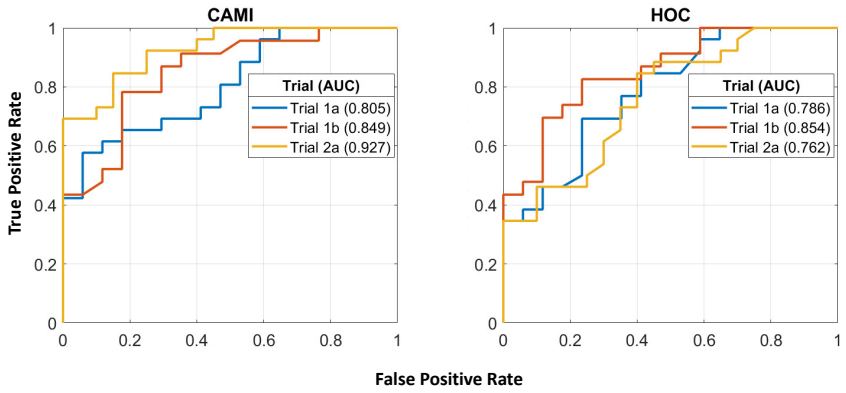


Figure S7. Correlations between the 10-fold cross-validated CAMI scores and HOC scores, showing strong correspondence between the two methods. A correlation value of r = 1 indicates perfect positive association, r = 0 indicates no association r = -1 indicates perfect negative association.

Using the 10-fold CAMI scores, the CAMI method's diagnostic classification ability was assessed using the same receiver operating characteristic (ROC) and support vector machine (SVM) approaches reported in the main text. This analysis revealed a slight improvement in the classification ability (see Figure S8), though the variability of the SVM accuracy increased considerably due to the small number of samples left for the test sets. In particular, the SVM accuracy of the CAMI scores ranged between 75% and 100% in the 10-fold scheme (SD = 12.2%), while the range was between 84.6% and 92.1% in the 3-fold scheme (SD = 3.6%). Similarly, the variability of the SVM accuracy of the HOC scores increased in the 10-fold scheme. Namely, while the SVM accuracy of the HOC scores ranged between 50% and 100% in the 10-fold scheme (SD = 15.8%), the range was between 61.5% and 84.6% in the 3-fold scheme (SD = 9.6%).

(A)



(B)

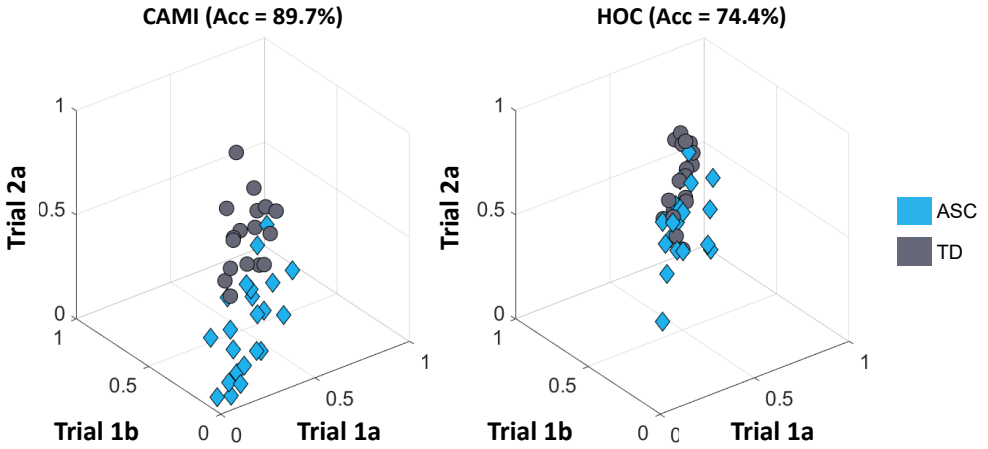


Figure S8. Diagnostic classification ability of the CAMI and HOC methods (CAMI scores obtained using 10-fold cross-validation).

(a) Receiving Operating Characteristic (ROC) curves: true positive rate versus false positive rate as classification threshold is varied. The Area Under the Curve (AUC) indicates the diagnostic ability of the method (blue lines for HOC, orange lines for CAMI method) in each of the three trials (best possible AUC is 1, meaning zero false positives and 100% true positives).

(b) 3D plots of the CAMI scores and HOC scores in which Trial 1a, Trial 1b, and Trial 2 scores correspond to the respective axes. Each marker represents one subject, and the reported accuracy (Acc) corresponds to average classification accuracy in 10-fold cross-validation of a linear SVM classifier (best possible Acc is 1, meaning 100% accuracy).

Supplemental References

[1] R. Poppe, S. van der Zee, D. K. J. Heylen and P. J. Taylor, AMAB: Automatic measurement and analysis of body motion, *Beh. Res. Methods*, 46(3), 625-633 (2013).

[2] T. Arici, S. Celebi, A. S. Aydin and T. T. Temiz, Robust gesture recognition using feature pre-processing and weighted dynamic time warping, *Multimedia Tools and App.*, 72(3), 3045-3062 (2014).

[3] M. Zanfir, M. Leordeanu and C. Sminchisescu, The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection, *Proc. of the IEEE Int. Conf. on Comp. Vision*, (2013), doi: 10.1109/ICCV.2013.342.

[4] E. Delaherche, S. Boucenna, K. Karp, S. Michelet, C. Achard and M. Chetouani, "Social coordination assessment: Distinguishing between shape and timing" in Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, F. Schwenker, S. Scherer, L. P. Morency, Eds., (Springer, Berlin, Heidelberg, 2012), doi: 10.1007/978-3-642-37081-6_2.

[5] S. Ferguson, E. Schubert and C. J. Stevens, Dynamic dance warping: Using dynamic time warping to compare dance movement performed under different conditions, Int. Conf. on Movement and Computing, 94-99 (2014), doi: 10.1145/2617995.2618012.

[6] H. Sakoe, and S. Chiba, Dynamic programming algorithm optimization for spoken word, *IEEE Transac. on Acoustics, Speech, and Signal Proc*., 26(1), 43-49 (1978).

[7] S. Celebi, A. S. Aydin, T. T. Temiz and T. Arici, Gesture recognition using skeleton data with weighted dynamic time warping, *Int. Conf. on Computer Vision Theory and App.*, (2013), doi: 10.5220/0004217606200625.

[8] D. Folgado, M. Barandas, R. Matias, R. Martins, M. Carvalho and H. Gamboa, Time alignment measurement for time series, *Pattern Recognition*, 81, 268-279 (2018).