Archival Report

Biological Psychiatry: CNNI

# Computerized Assessment of Motor Imitation as a Scalable Method for Distinguishing Children With Autism

Bahar Tunçgenç, Carolina Pacheco, Rebecca Rochowiak, Rosemary Nicholas, Sundararaman Rengarajan, Erin Zou, Brice Messenger, René Vidal, and Stewart H. Mostofsky

## ABSTRACT

**BACKGROUND:** Imitation deficits are prevalent in autism spectrum conditions (ASCs) and are associated with core autistic traits. Imitating others' actions is central to the development of social skills in typically developing populations, as it facilitates social learning and bond formation. We present a Computerized Assessment of Motor Imitation (CAMI) using a brief (1-min), highly engaging video game task.

**METHODS:** Using Kinect Xbox motion tracking technology, we recorded 48 children (27 with ASCs, 21 typically developing) as they imitated a model's dance movements. We implemented an algorithm based on metric learning and dynamic time warping that automatically detects and evaluates the important joints and returns a score considering spatial position and timing differences between the child and the model. To establish construct validity and reliability, we compared imitation performance measured by the CAMI method to the more traditional human observation coding (HOC) method across repeated trials and two different movement sequences.

**RESULTS:** Results revealed poorer imitation in children with ASCs than in typically developing children ($ps < .005$), with poorer imitation being associated with increased core autism symptoms. While strong correlations between the CAMI and HOC methods ($rs = .69–.87$) confirmed the CAMI's construct validity, CAMI scores classified the children into diagnostic groups better than the HOC scores ($accuracy_{CAMI} = 87.2\%$, $accuracy_{HOC} = 74.4\%$). Finally, by comparing repeated movement trials, we demonstrated high test-retest reliability of CAMI ($rs = .73–.86$).

**CONCLUSIONS:** Findings support the CAMI as an objective, highly scalable, directly interpretable method for assessing motor imitation differences, providing a promising biomarker for defining biologically meaningful ASC subtypes and guiding intervention.

https://doi.org/10.1016/j.bpsc.2020.09.001

Imitating others' actions is crucial for social bond formation and learning (1–3), with atypical imitation indicating social-communicative impairments in autism spectrum conditions (ASCs) (4–8). The current standard in imitation assessment is manual human observation coding (HOC), which is subjective and time-consuming and requires intensive coder training. These drawbacks render HOC impractical for use in clinics and home settings. Automatic assessment of imitation is challenging because human motion data are highly heterogeneous (e.g., range of movements is virtually unlimited) and high dimensional (motion data involve spatial and temporal aspects), and human supervision (e.g., expert knowledge) is limited and error prone. Addressing these issues, we present a Computerized Assessment of Motor Imitation (CAMI) that can improve diagnosis and treatment efforts by providing an objective, continuous, and scalable score of imitation performance.

Examining imitation performance with HOC methods requires identifying individual steps involved in an action, the action's style, the order of occurrence, repetitions, and the end goal, if it exists (9–12). Participants receive an ordinal score depending on the correct actions and errors made. Thus, the accuracy and precision (i.e., sampling frequency) of HOC is restricted by human subjectivity. First, what constitutes a "good enough" resemblance to the target action is at the human observers' discretion. Moreover, the defined action categories may likely miss preliminary forms of that action (e.g., flexing the fingers wide open without moving the hand when assessing the action of waving). Finally, assessments are subjective even within agreed-upon standards, an issue often circumvented by seeking high interrater reliability from multiple coders. Although this workaround alleviates subjectivity, it adds to coder training and assessment time. As such, the HOC method is largely confined to research settings and is not conducive to practical use as a diagnostic or treatment tool.

Prior attempts to develop automated methods for assessing imitation performance have focused on determining the match between a participant's movements and those of a template. The two most commonly used methods are rule-based algorithms (13,14) and algorithms based on dynamic time warping



ISSN: 2451-9022

(DTW) (15,16). Similar to HOC, rule-based algorithms require the researchers to manually define a set of rules. How well the participants meet these rules is automatically assessed by the algorithm. Despite its demonstrated utility in robot-mediated therapy settings with children with ASCs (13,14), rule-based methods have very limited generalizability, as they require a priori human input for selecting the rules specific to the gestures under study.

In contrast, DTW-based methods assess the spatial similarity between two time series after correcting for discrepancies in the temporal dimension (17) without requiring human input. DTW-based approaches have been widely used in gesture recognition tasks, in which a decision about which gesture the participant performed is outputted based on the similarity between the participant and the template (18–21). Existing DTW-based imitation assessment approaches define a metric that is either dichotomous (imitated vs. not) (15) or categorical (good vs. bad performance) (16), thereby not utilizing the continuous distance metric obtained from DTW. These approaches can only capture relatively large variations in imitation performance because these approaches have categorical outputs, an issue that becomes even more prominent in clinical populations such as ASCs that display high behavioral variability. Moreover, given the importance of timing in social coordination (22) and in characterizing autism-specific imitation impairments (23,24), a valid imitation assessment system must also consider temporal differences.

Machine learning techniques that learn motor patterns and classify individuals into diagnostic groups is another popular approach (25–27). Yet, prior studies did not directly assess imitation ability even when some tasks involved imitation (25,26). Therefore, it is unclear if the observed differences represent general motor abnormalities or specific imitation impairments. This lack of specificity restricts the use of these methods for intervention purposes.

Characterizing and addressing imitation impairments is crucial because imitation plays an important role in social bond formation and learning (1–3), joint attention (12), children's play initiation (28), and social affiliation and prosocial behaviors (29,30). Extant research shows that as compared with their typically developing (TD) peers, imitation in children with ASCs is less frequent, less precise, and more delayed (9,10,31,32). These imitation deficits are more pronounced when the actions appear meaningless or lack an obvious end goal (5,6,9,11,24). Impaired imitation is associated with poorer social-communicative functioning in children with ASCs, as demonstrated in social responsiveness, social attention, engagement in joint play, and social reciprocity (33–35).

An automated method that 1) specifically measures imitation performance, 2) does not require manual feature selection, 3) generalizes to a range of movement types, 4) provides continuous scores, and 5) integrates spatial and timing differences would significantly improve diagnosis efforts and robot-mediated and other social-communication interventions for autism.

The CAMI method that we developed uses 3-dimensional motion data obtained from sensorless Kinect Xbox cameras (Microsoft, Redmond, WA). Using DTW and metric learning techniques, CAMI considers differences in both motion trajectories and timing differences. We applied CAMI to a dataset comprising 48 children (27 with ASCs, 21 TD) as they imitated the dance-like movements of a video model. In this article, we report on the construct validity of the CAMI method, assessed by comparing the children's CAMI scores with their imitation scores obtained by the HOC method. We also established the test-retest reliability of the CAMI by comparing the children's scores across repeated imitation trials. We demonstrate how well imitation scores from the CAMI versus HOC methods classify the children into diagnostic categories. Finally, we present the CAMI's clinical significance by examining imitation performance in the ASC and TD groups and its association with core autism symptoms.

We hypothesized that while CAMI scores would highly correlate with HOC scores and show high test-retest reliability, the CAMI would outperform HOC when used for distinguishing the diagnostic groups. Further, we expected that CAMI scores would yield clinically meaningful results by revealing poorer imitation in children with ASCs as compared with TD children and showing strong associations with core autism symptoms.

## METHODS AND MATERIALS

### Participants

The data reported here were collected as part of a wider-scale study examining imitation skills in autism. Our participants were 48 children (27 with ASC, 21 TD) 8 to 12 years of age.

Autism diagnosis was based on DSM-5 criteria and was confirmed on site by research-reliable assessors using the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2), the Autism Diagnostic Interview–Revised, and the parent-report of the Social Responsiveness Scale, Second Edition (SRS-2). To be included in the study, children needed a full-scale IQ score ≥80 or at least one index score ≥80 (verbal comprehension; visual, spatial, or fluid reasoning index) on the Wechsler Intelligence Scale for Children–Fifth Edition. For all participants, ADOS-2 module 3 was used. In addition, to account for autism-associated differences in general motor abilities, we used the Movement Assessment Battery for Children, Second Edition. Descriptive statistics of participant characteristics can be found in Table 1. See the Supplement for full inclusion/exclusion criteria.

Ethics approval was received from the Johns Hopkins University School of Medicine Institutional Review Board prior to study commencement. Written informed consent was obtained from all participants' legal guardians, and verbal assent from all children. All recruitment took place through contacts with local schools and community events. Participants were invited to the Center for Neurodevelopmental and Imagine Research at the Kennedy Krieger Institute for 2-day visits and received $100 compensation for their time.

### Procedure

Children took part in an imitation task comprising 14 trials presented at varying movement speeds. To avoid any confound of changing movement speeds, in this study, we report only on 4 trials presented at 100% speed: the first two trials (trials 1a and 2a) and the last two trials (trials 1b and 2b). The last two trials were repetitions of the first two trials. These trials were of two separate movement sequences (sequence

**Table 1. Participant Characteristics**

| | ASC Group | | TD Group | | |
|---|---|---|---|---|---|
| | Mean (SD) or *n* | Range | Mean (SD) or *n* | Range | Test Statistic (ASC vs. TD Group) |
| Chronological Age, Years | 10.34 (1.42) | 8.03–12.83 | 10.41 (1.26) | 8.55–12.73 | $t_{46} = -0.19$, $p > .05$ |
| SRS-2 Total Score | 75.73 (7.40) | 60–87 | 44.81 (4.04) | 39–54 | $t_{40} = 15.33$, $p < .0001$ |
| ADOS-2 Total Score | 15.54 (4.34) | 8–27 | – | – | – |
| WISC-V Full Scale IQ | 98.41 (15.62) | 70–130 | 109.86 (12.06) | 94–143 | $t_{46} = -2.78$, $p = .008$ |
| Boys/Girls | 24/3 | – | 18/3 | – | $\chi^2_{1,48} = 0.11$, $p > .05$ |

ADOS-2, Autism Diagnostic Observation Scale, Second Edition; ASC, autism spectrum condition; SRS-2, Social Responsiveness Scale, Second Edition; TD, typically developing; WISC-V, Wechsler Intelligence Scale for Children, Fifth Edition.

1 = trials 1a and 1b, sequence 2 = trials 2a and 2b). The sequences comprised 14 to 18 individual movement types, which were relatively unfamiliar (e.g., moving arms up and down like a puppeteer), did not have an end goal, and required moving multiple limbs simultaneously. The choice of these movement sequences was based on prior research showing particular difficulties in ASCs with these types of movements (5,6,8,9,11).

The stimulus video was displayed on a large TV screen and depicted dance-like whole-body movements of a young woman without any background music/sound. The children's movements were recorded using two Kinect Xbox cameras at 30 frames/s, one located in front of the child and one at the back. Because the Kinect Xbox records depth data, no sensors or special clothing were needed for this data collection. For more information about the study setup, see the Supplement.

The session began with a brief training phase, familiarizing the participants to the kinds of movements that they would perform and how much to move their bodies. All participants were instructed to perform whole-body movements and to try their best to copy the model.

## Data Coding

**Calculation of CAMI Scores.** The x-y-z coordinates of 20 joints were extracted from the children's depth recordings using iPi Motion Capture Software (IpiSoft, Moscow, Russia). Children's motion data were compared with the "gold standard," defined here as the motion data of the video model imitating herself. Imitation scores for each child were obtained following the steps outlined below. The details of the CAMI method and equations used can be found in Computerised Assessment of Motor Imitation (CAMI) Algorithm in the Supplement.

1. Preprocessing. The child's and the gold standard's motion data are translated by locating their hips' positions at the origin. The child's limb lengths are normalized to the gold standard's skeleton, and the child's spatial orientation in the first frame is adjusted to match the gold standard.
2. Automatic joint importance estimation. Using the gold standard data, the relative contributions of each joint for each movement type are computed based on the amount of displacement observed. Joints that were displaced more in the gold standard data for a given movement type are considered to contribute more to the movement and hence affected the imitation score more than joints that stayed static.
3. Computation of the distance feature. Using DTW (17), the child's time course is aligned to the model's time course for the entirety of the sequence by finding a time warp that minimizes the Euclidean distance between them. The DTW distances of each movement type are calculated considering the relative importance of each joint as computed in step 2. The distances for the movement types are then averaged to make up the child's total DTW distance (dist), which is then transformed into a distance score ($s_{dist}$).
4. Computation of the time features. Using the DTW warping path information, time asynchrony features are computed for the entire sequence (36): the duration that children were delayed with respect to the model ($t_{delay}$) and the duration that children performed the movements in advance of the model ($t_{adv}$).
5. Computation of the CAMI score. Using metric learning techniques, the 3 variables ($s_{dist}$, $t_{delay}$, and $t_{adv}$) are linearly combined to make up the child's imitation score. The weights used for this linear combination are determined in a data-driven manner using a 3-fold cross-validation technique. In this technique, first, the dataset is split into 3 nonoverlapping groups with equal proportions of children with ASCs and TD children in each group. Then, two of these groups (i.e., training set) are used to learn the weights in a way that maximizes the average correlation between the CAMI and HOC across the trials of the training set. Using the learned parameters, the CAMI scores of the third group (i.e., the test set) are calculated. Using cross-validation ensures that children's CAMI scores are calculated completely independently from and without reference to their HOC scores. The same procedure is repeated by assigning a new group as the test set until the CAMI scores are obtained for all three groups. The formula used for learning the weights, and the parameter values can be found in Parameter Learning in the Supplement.

Regarding the number of cross-validation folds, studies have shown that too few folds can lead to biased estimators, and too many folds generate high variance in the estimations (37). Hence, we repeated the analyses using 10 folds, which replicated the same findings (see Supplemental Results).

Because variability was considerably larger in the 10-fold scheme (12.2%) as compared with the 3-fold scheme (3.6%), we are reporting the findings from the 3-fold scheme in the main text.

### Calculation of HOC Scores

To establish construct validity of the CAMI method, we analyzed 3 trials (trials 1a, 1b, and 2a) using the more traditional HOC method. At least 40% of the videos within each trial, evenly split across diagnostic groups, were reliability-coded by 2 diagnosis-blind coders (all $K > .92$, $ps < .001$). No HOC was done for trial 2b videos so that this trial could be used as the replication dataset.

Our HOC scheme identified the components of all movement types within a sequence (e.g., bring right arm to the right), the style of the movements (e.g., twirl arm, right/left side), and the number of repetitions. Children's total HOC score was the sum of positive items ($s_{pos}$) and negative items ($s_{neg}$) for each movement type, divided by the maximum possible score for that movement type. ($s_{pos}$) comprised scores given to components successfully completed (score of +1). ($s_{neg}$) comprised scores given to movements performed on the reverse side (score of −0.5) and to movements that were repeated more times than demonstrated by the model (score of −1). Consequently, the children could receive a score within the range of 0 to 166 for sequence 1, which had 176 components, and 0 to 202 for sequence 2, which had 216 components. These scores were then normalized to a range of 0 to 1, where 1 indicates perfect imitation and 0 indicates worst imitation [see Human Observation Coding (HOC) Scheme in the Supplement].

## RESULTS

To enable replications and use by future research, we provide the learned parameters of the CAMI in the Supplemental Results. While developing the CAMI method, we proposed that it would have 3 main advantages to alternative methods: 1) considering temporal, in addition to spatial, differences in imitation performance; 2) assessing imitation ability with high sensitivity by yielding continuous, rather than discrete, scores; and 3) automatically detecting which joints are important for different movement types without human input. Beyond theoretical plausibility of these arguments, we conducted rigorous experiments, which empirically confirmed that these properties did indeed improve the CAMI's performance (see Supplemental Results).

### Construct Validity and Test-Retest Reliability of the CAMI

To establish the CAMI's construct validity, we examined its correlation with the scores obtained from HOC in three trials of movement data. The results revealed strong positive correlations between the two methods for all 3 trials (trial 1a: $r_{43} = .82$, $p < .0001$; trial 1b: $r_{40} = .87$, $p < .0001$; trial 2a: $r_{46} = .69$, $p < .0001$) (Figure 1A). Notably, the correlation between the two methods was lowest for trial 2a. It is worth highlighting here that the CAMI scores are calculated using a 3-fold cross-validation method, which means that children's CAMI scores were calculated independently from their HOC scores. Further supporting this point, when the same correlation tests were run between HOC scores and the distance output of DTW, which is completely unsupervised by HOC, we still observed strong correlations between the two variables (trial 1a: $r_{43} = -.78$, $p < .0001$; trial 1b: $r_{40} = -.82$, $p < .0001$; trial 2a: $r_{46} = -.70$, $p < .0001$), such that increased spatial difference between the child and the model was correlated with worse HOC scores.

We assessed the CAMI's test-retest reliability by comparing performance scores between repetitions trials, comparing trial 1a with trial 1b and trial 2a with trial 2b with Pearson's correlation tests. The results revealed excellent test-retest reliability (trials 1a and 1b: $r_{37} = .86$, $p < .0001$; trials 2a and 2b: $r_{36} = .73$, $p < .0001$).

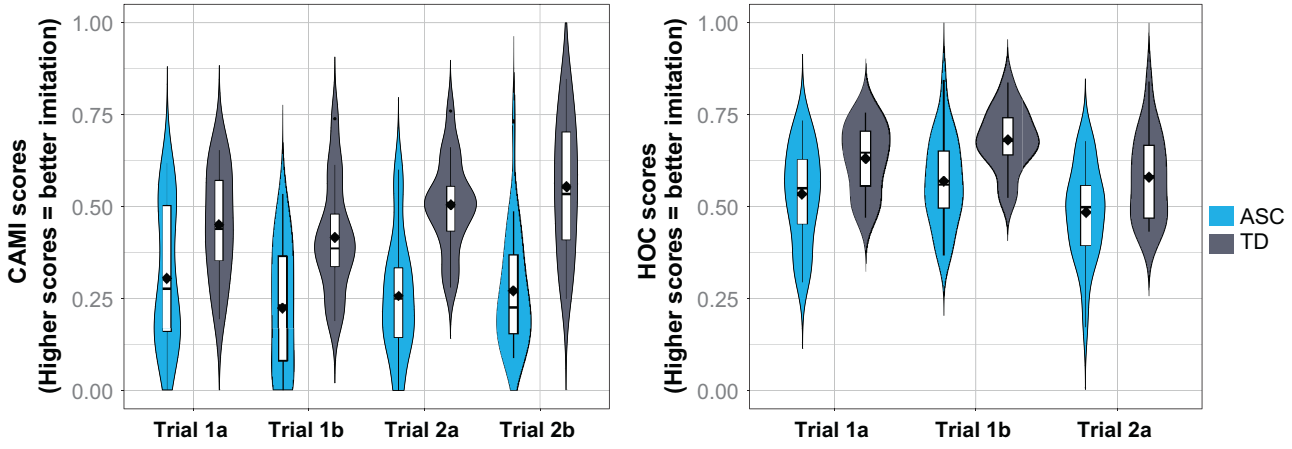


**Figure 1.** Imitation performance per diagnostic group (blue = autism spectrum conditions [ASCs], gray = typically developing [TD]) per trial according to Computerized Assessment of Motor Imitation (CAMI) (left) and human observation coding (HOC) scores (right) with box plots embedded within violin plots. In the box plots, horizontal lines indicate medians, boxes indicate data within the 25th to 75th percentiles, and whiskers indicate data within the 5th to 95th percentiles.

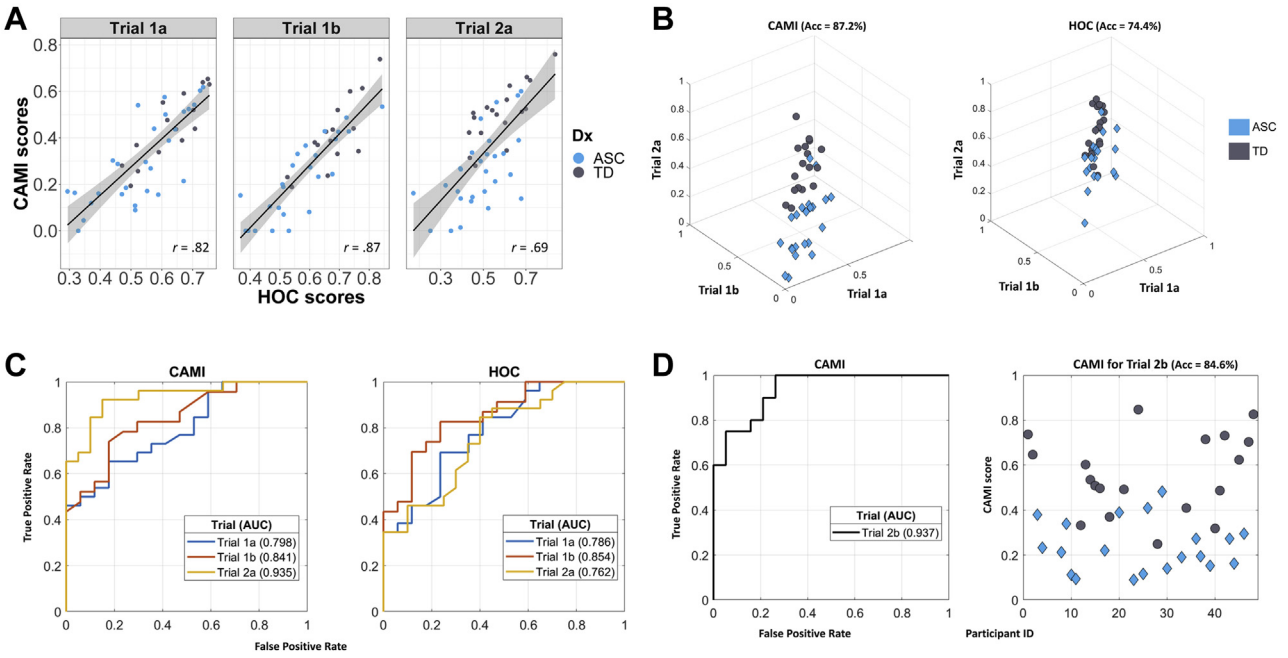


**Figure 2.** Comparisons between the Computerized Assessment of Motor Imitation (CAMI) and human observation coding (HOC) methods using motion data of 4 imitation trials from a sample of 48 children (27 with autism spectrum conditions [ASCs], 21 typically developing [TD]). **(A)** Correlations between the CAMI scores and HOC scores in 3 trials, showing strong correspondence between the two methods. An $r = 1$ indicates perfect positive association, $r = 0$ indicates no association and $r = -1$ indicates perfect negative association ($ps < .0001$). **(B)** Three-dimensional plots of the CAMI and HOC scores in which scores from trials 1a, 1b, and 2a correspond to the respective axes. Each marker represents 1 subject, and the reported accuracy (Acc) corresponds to average classification accuracy in 3-fold cross-validation of a linear support vector machine (SVM) classifier (best possible Acc is 100%, meaning all participants categorized to diagnostic groups accurately). **(C)** Receiver-operating characteristic curves: true positive rate vs. false positive rate as classification threshold is varied. The area under the curve (AUC) indicates the diagnostic ability of the method (left panel for CAMI, right panel for HOC) in each of the 3 trials (best possible AUC is 1, meaning 0 false positives and 100% true positives). **(D)** Receiver-operating characteristic curve (left) and CAMI scores (right) of trial 2b only. Because this trial did not have any HOC scores, its CAMI scores were computed based on parameters learnt from the other 3 trials, complying with the splits used for 3-fold cross-validation. The area under the receiver-operating characteristic curve (0.937) and SVM accuracy value (84.6%) demonstrate the diagnostic classification ability of CAMI scores with this single trial. Dx, diagnosis.

## Diagnostic Classification Ability of the CAMI

We assessed how well imitation scores obtained from the CAMI and HOC methods would classify children into diagnostic groups in two ways: 1) by training a standard machine learning algorithm (linear support vector machines [SVMs]) to classify the subjects into diagnostic groups using their imitation scores as the sole features; and 2) by computing the receiver-operating characteristic curve of each trial, with larger areas under the curve indicating better discriminative ability. Notably, the features used in SVMs carry no prior information about the children's diagnosis status; the only feature used for classification was the diagnosis-blind imitation scores.

For all 3 trials that had both CAMI and HOC scores, the discriminative ability of the CAMI was either comparable to or better than the discriminative ability of HOC. The SVM method showed that when the participants are characterized by their imitation performance across the trials, the diagnostic groups (ASCs vs. TD) are more vividly separated by CAMI scores as compared with HOC scores (Figure 1B). This visible trend is supported by the higher average classification accuracy obtained by a linear SVM classifier trained in a 3-fold cross validation scheme ($accuracy_{CAMI} = 87.2\%$, $accuracy_{HOC} = 74.4\%$).

The area under the curve was comparable between the CAMI and HOC methods for trials 1a and 1b, while it was considerably higher for CAMI scores in trial 2a (Figure 1C). Considered together with the lower correlation between CAMI and HOC in trial 2a (Figure 1A), this finding attests to CAMI's validity and superior diagnostic classification ability.

Finally, we replicated these findings in a single, 1-minute imitation trial (trial 2b), which had only CAMI scores and no HOC scores. Figure 1D shows the classification accuracy (84.6%) and area under the curve of the CAMI based only on trial 2b scores. Overall, these findings show that the CAMI outperforms HOC in distinguishing children into diagnostic groups.

## Clinical Relevance of the CAMI

To confirm that the imitation ability assessed by the CAMI is relevant for a clinical autism sample, we examined the hypotheses that poorer imitation would be observed in children with ASCs than in TD children, and that imitation deficits would be associated with increased autism symptoms.

We conducted a mixed analysis of variance with diagnosis (ASCs vs. TD), assessment method (CAMI vs. HOC), age, IQ, and motor abilities (scores from the Movement Assessment Battery for Children–Second Edition) as the independent variables, and imitation score as the dependent variable. One TD

**Table 2. Correlation Between Imitation Ability and Core Autism Symptoms**

| | SRS-2 (Parent Report) | | | ADOS-2 (in the ASC Group Only) | | |
|---|---|---|---|---|---|---|
| Imitation Ability | SA Subscale | RRB Subscale | Total Score | SA Subscale | RRB Subscale | Total Score |
| CAMI Method | $r_{40} = -.48^{b}$ | $r_{40} = -.52^{c}$ | $r_{40} = -.50^{b}$ | $r_{24} = -.17$ | $r_{24} = -.23^{b}$ | $r_{24} = -.34^{a}$ |
| HOC Method | $r_{40} = -.42^{b}$ | $r_{40} = -.45^{b}$ | $r_{40} = -.43^{b}$ | $r_{24} = -39^{a}$ | $r_{24} = -.21$ | $r_{24} = -.42^{b}$ |

$r = 1$ indicates perfect positive association, $r = 0$ indicates no association, and $r = -1$ indicates perfect negative association. Higher scores in the SRS-2 and ADOS-2 indicate greater autism severity.

ADOS-2, Autism Diagnostic Observation Scale, Second Edition; ASC, autism spectrum condition; CAMI, Computerized Assessment of Motor Imitation; HOC, human observation coding; RRB, restricted interests and repetitive behavior; SA, social affect; SRS-2, Social Responsiveness Scale, Second Edition.

[a] $p < .10$.

[b] $p < .05$.

[c] $p < .001$.

child and 2 children with ASCs were dropped from analysis owing to violation of normality assumptions (±2 SDs from the mean); including these children did not change the findings. These statistical analyses were done using open-source R software version 3.6.2 (R Foundation for Statistical Computing, Vienna, Austria).

The results revealed significant main effects of diagnosis ($F_{1,23} = 13.41$, $p = .001$) and assessment method ($F_{1,25} = 221.69$, $p < .0001$) as well as significant interaction effects of diagnosis × trial ($F_{2,92} = 3.56$, $p = .03$), diagnosis × method ($F_{1,25} = 14.80$, $p = .0007$), and trial × method ($F_{1,92} = 10.55$, $p = .0001$). No other variable had a significant effect on imitation scores. Owing to its relevance for our hypothesis, we further examined the diagnosis × trial interaction with Bonferroni-corrected pairwise tests. We found that within each trial and for both CAMI and HOC scores, children with ASCs imitated more poorly than TD children (all $p$s $< .0001$) (Figure 2). That no ceiling effects were observed in either group indicates that, as reported in the prior literature (5,6,8,9,11), the types of movements included in these sequences were challenging for both the ASC and TD groups.

To examine the associations between imitation performance and core autism symptoms, we created composite scores by averaging the children's scores in 3 trials (trials 1a, 1b, and 2a). Core autism symptoms were measured by 1) parental reports of the SRS-2 and 2) the ADOS-2 administered to children with ASCs. Better imitation ability, as measured by both the CAMI and HOC methods, was moderately and statistically significantly correlated with lower scores on the subscales of the SRS-2 and the total SRS-2 scores (Table 2). Correlations between the imitation scores and ADOS-2 scores were less strong, with the total scores reaching or approaching significance. The decreased association of CAMI scores with ADOS-2 scores likely stemmed from insufficient power because the ADOS-2 was administered only to the ASC group, while the SRS-2 was administered to all participants. These findings support the clinical relevance of CAMI by revealing significant links between the CAMI-assessed imitation deficits and core autism symptoms.

## DISCUSSION

In this study, we developed a method called CAMI, which uses an automated, DTW-based algorithm, and presented its successful application to a clinical autism population to examine imitation deficits. Strong correspondence of the CAMI with the standard HOC method confirmed our method's construct validity. Applying the CAMI on two sets of repeated imitation trials involving two different movement sequences, we established the CAMI's test-retest reliability. Further, the findings revealed that imitation ability as assessed by CAMI scores can distinguish children's clinical diagnosis (ASCs vs. TD) better than HOC scores. Clinical relevance of the CAMI has been further confirmed with findings of CAMI-assessed poorer imitation in children with ASC than in TD children, and a strong link between imitation deficits and core autism symptom severity.

The CAMI addresses the outstanding issues with automatic assessment of human motion and imitation. The issues of heterogeneity (i.e., range of movements being virtually unlimited) and requirement for high sensitivity to detect nuances are addressed by using a continuous instead of a discrete output. The issues of high dimensionality (i.e., involving spatial and temporal aspects of movements) and limited human supervision (i.e., lacking expert knowledge on importance of movement elements) are addressed by imposing a structure in the model that reduces the number of learnable parameters based on guidance from expert-based observations (i.e., HOC scores). Using expert knowledge to guide the features (i.e., dist, $t_{adv}$, and $t_{delay}$) improved CAMI's interpretability, while deviance from HOC as a result of automatized learning processes improved CAMI's diagnostic discriminative ability. Finally, automatic detection of important joints enables combining high-dimensional data in a meaningful way for other movements, improving the CAMI's scalability.

One advantage of the CAMI is that unlike other automated methods that classify children into diagnostic groups based on broad differences in movement patterns (25–27), it specifically measures imitation ability and provides an interpretable score indicating how well the children performed with respect to a model. Assessing imitation ability with a sensitive, automatic, and objective method is important because there is robust evidence that imitation crucially impacts social bonding, learning, communication, and interaction throughout development (1,2,3,30). Because our method targets imitation ability in particular, it can be used to detect deficiencies from at least school age onward and to track performance during interventions designed to improve social-communicative function through imitation training.

Contrary to previous methods providing only dichotomous or categorical scores for imitation performance (15,16), CAMI produces a fine-grained, continuous score within a range of 0 (worst imitation) to 1 (best imitation). Continuous scores allow for capturing minute differences in imitation ability, which is especially important for populations with high variability such as children with ASCs. Moreover, the CAMI considers timing differences in addition to spatial differences. Given known deficits in coordinating the timing of actions in ASCs (24,31), timing measures may importantly improve the assessment of autism-specific imitation impairments.

Using an SVM approach with 3-fold cross-validation, we demonstrated that CAMI scores designated children into diagnostic groups better than HOC scores. The CAMI method outperformed the HOC method in trial 2a, which was the trial with lowest correlation between the two methods. Further, applied to another trial without HOC scores (trial 2b), CAMI scores from a single, 1-minute trial distinguished the children into diagnostic groups with 84.6% success. Altogether, our findings show that as compared with HOC, the CAMI is more sensitive in detecting autism-associated differences in imitation performance.

Given the heterogeneity of the behavioral phenotypes in the autism spectrum, the sample size of 27 children with ASCs and 21 TD children can be considered relatively small. While such inherent limitations of a small sample size should be considered while interpreting the findings, it is important to clarify that the CAMI machine learning approach does not suffer from a sample size problem in either the calculation of the CAMI scores or SVM classification. First, because estimating the minimum sample size needed by canonical correlation analysis (a general case of the method we used to learn the CAMI scores) is nontrivial, studies suggest the "1-in-10 rule," which states that 10 samples per variable is enough to estimate the parameters (38,39). In our case, when we maximize the correlation, we are working with 3 variables; thus, 30 samples should suffice. In the 3-fold cross-validation scheme, we are using approximately 90 samples (two-thirds of subjects, 3 trials) to estimate 3 variables. Second, for binary max-margin linear classifiers such as the SVMs we used, Raudys (40) provided a formula to estimate the mean expected classification error in terms of the number of parameters, distance between classes, and sample size. Applying equation 12 from Raudys (40) to our problem, in which p = 3 (number of parameters), delta = 3.83 (the distance between normalized class centers), and $n$ = 13 (the approximate number of samples per class for training in the 3-fold cross-validation scheme), we obtained a mean expected classification error of 5.9%. Given that the minimum possible mean expected classification error for this problem (i.e., if we had infinite samples) would be 2.8%, our error rate can be considered sufficiently good. Notably, the parameter-to-sample-size ratio used here is higher than previous applications of machine learning to distinguish motor patterns in autism. For example, Li *et al.* (26) trained a model with 40 parameters using data from 30 subjects, and Crippa *et al.* (27) trained a model with 7 parameters using data from 30 subjects.

Future research is needed to improve the scalability of CAMI. At present, this method uses data obtained from Kinect Xbox depth cameras, which, owing to imperfections of the motion tracking technology, require some manual processing that can be time-consuming. Future research should explore the use of this method on 2-dimensional data obtained from high-resolution cameras. Moreover, in order for the CAMI to be used widely as a clinical assessment tool, we need to establish norm standardization with larger datasets, including younger age groups and varied demographics. Relatedly, given the relatively small sample size of this study for the highly heterogeneous autism population, it is crucial that the current findings be replicated in future research; administering the ADOS-2 to the entire population would be informative in future replication attempts.

The CAMI method presented here is a major step forward in examining motor imitation automatically without requiring extensive human input or coder training. This method provides an objective, continuous, highly scalable, and directly interpretable score. As such, the CAMI can be used in clinics and home settings to assess imitation ability, to help inform diagnostic decision making based on the children's imitation performance (e.g., ASCs vs. non-ASCs), and to advance biomarker-based interventions for improving social-communicative functioning through imitation-based strategies.

## ACKNOWLEDGMENTS AND DISCLOSURES

Recruitment for this study was supported by National Institutes of Health Grant No. R01 MH106564-02, for which SHM is a co-investigator, while algorithm development was supported by National Institutes of Health Grant No. R01 HD87133-01, for which RV is a co-investigator.

The authors report no biomedical financial interests or potential conflicts of interest.

## ARTICLE INFORMATION

From the Center for Neurodevelopment and Imaging Research (BT, RR, EZ, BM, SHM), Kennedy Krieger Institute; Mathematical Institute for Data Science (CP, RV) and Department of Biomedical Engineering (CP, RV), Johns Hopkins University; and the Department of Neurology (SHM) and Department of Psychiatry and Behavioral Sciences (SHM), Johns Hopkins University School of Medicine, Baltimore, Maryland; School of Psychology (BT) and Sir Peter Mansfield Imaging Centre (RN), University of Nottingham, Nottingham, United Kingdom; and the Joint Doctoral Program in Language and Communicative Disorders (SR), San Diego State University and University of California San Diego, San Diego, California.

Address correspondence to Bahar Tunçgenç, Ph.D., at bahartuncgenc@gmail.com.

Received Jun 18, 2020; revised and accepted Sep 2, 2020.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.bpsc.2020.09.001.

## REFERENCES

1. Prinz W (2002): Experimental approaches to imitation. In: Melzoff AN, Prinz W, editors. The Imitative Mind: Development, Evolution and Brain Bases. Cambridge, United Kingdom: Cambridge University Press, 143–162.
2. Heyes C (2013): What can imitation do for cooperation? In: Sterelny K, Joyce R, Calcott B, Fraser B, editors. Cooperation and Its Evolution. Life and Mind: Philosophical Issues in Biology and Psychology Cambridge, MA: MIT Press, 313–331.
3. Nadel J (2006): Does imitation matter to children with autism? In: Rogers SJ, Williams JHG, editors. Imitation and the Social Mind: Autism and Typical Development New York, NY: Guillford Press, 118–137.
4. Edwards LA (2014): A meta-analysis of imitation abilities in individuals with autism spectrum disorders. Autism Res 7:363–380.

5. Gowen E (2012): Imitation in autism: Why action kinematics matter. Front Int Neurosci 6:117.
6. MacNeil LK, Mostofsky SH (2012): Specificity of dyspraxia in children with autism. Neuropsychology 26:165–171.
7. Nebel MB, Eloyan A, Nettles CA, Sweeney KL, Ament K, Ward RE, *et al.* (2016): Intrinsic visual-motor synchrony correlates with social deficits in autism. Biol Psychiatry 79:633–641.
8. McAuliffe D, Pillai AS, Tiedemann A, Mostofsky SH, Ewen JB (2017): Dyspraxia in ASD: Impaired coordination of movement elements. Autism Res 10:648–652.
9. Hobson PR, Hobson JA (2008): Dissociable aspects of imitation: A study in autism. J Exp Child Psychol 101:170–185.
10. Vivanti G, Nadig A, Ozonoff S, Rogers SJ (2008): What do children with autism attend to during imitation tasks? J Exp Child Psychol 101:186–205.
11. Marsh L, Pearson A, Ropar D, Hamilton A (2013): Children with autism do not overimitate. Curr Biol 23:R266–R268.
12. Carpenter M, Tomasello M, Savage-Rumbaugh S (1995): Joint attention and imitative learning in children, chimpanzees, and enculturated chimpanzees. Soc Dev 4:217–237.
13. Zheng Z, Das S, Young EM, Swanson A, Warren Z, Sarkar N (2014): Autonomous robot-mediated imitation learning for children with autism. IEEE Int Conf Robot Autom 2014:2707–2712.
14. Zheng Z, Young EM, Swanson AR, Weitlauf AS, Warren ZE, Sarkar N (2016): Robot-mediated imitation skill training for children with autism. IEEE Trans Neural Syst Rehabil Eng 24:682–691.
15. Michelet S, Karp K, Delaherche E, Achard C, Chetouani M (2012): Automatic imitation assessment in interaction. In: Salah AA, Ruiz-del-Solar J, Meriçli C, Oudeyer PY, editors. HBU 2012: Human Behavior Understanding. Lecture Notes in Computer Science, vol. 7559. Berlin, Germany: Springer, 161–173.
16. Su CJ, Chiang CY, Huang JY (2014): Kinect-enabled home-based rehabilitation system using Dynamic Time Warping and fuzzy logic. Appl Soft Comput 22:652–666.
17. Sakoe H, Chiba S (1978): Dynamic programming algorithm optimization for spoken word. IEEE Trans Acoust Speech Signal Process 26:43–49.
18. Reyes M, Dominguez G, Escalera S (2011): Feature weighting in dynamic time warping for gesture recognitoin in depth data. IEEE Int Conf Comput Vis Workshops 2011:1182–1188.
19. Celebi S, Aydin AS, Temiz TT, Arici T (2013): Gesture recognition using skeleton data with weighted dynamic time warping. Proc Int Conf Comput Vis Theory Appl 1:620–625.
20. Ofli F, Chaudhry R, Kurillo G, Vidal R, Bajcsy R (2014): Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. J Vis Commun Image Represent 25:24–38.
21. Arici T, Celebi S, Aydin AS, Temiz TT (2014): Robust gesture recognition using feature pre-processing and weighted dynamic time warping. Multimed Tools Appl 72:3045–3062.
22. Delaherche E, Boucenna S, Karp K, Michelet S, Achard C, Chetouani M (2012): Social coordination assessment: Distinguishing between shape and timing. In: Schwenker F, Scherer S, Morency LP, editors. MPRSS: Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction. Berlin, Germany: Springer, 9–18.
23. de Marchena A, Eigsti IM (2010): Conversational gestures in autism spectrum disorders: Asynchrony but not decreased frequency. Autism Res 3:311–322.
24. Marsh KL, Isenhower RW, Richardson MJ, Helt M, Verbalis AD, Schmidt RC, Fein D (2013): Autism and social disconnection in interpersonal rocking. Front Int Neurosci 7:108.
25. Guha T, Yang Z, Grossman RB, Narayanan SS (2015): A computational study of expressive facial dynamics in children with autism. IEEE Trans Affect Comput 9:14–20.
26. Li B, Sharma A, Meng J, Purushwalkam S, Gowen E (2017): Applying machine learning to identify autistic adults using imitation: An exploratory study. PLoS One 12:e0182652.
27. Crippa A, Salvatore C, Perego P, Forti S, Nobile M, Molteni M, Castiglioni I (2015): Use of machine learning to identify children with autism and their motor abnormalities. J Autism Dev Disord 45:2146–2156.
28. Fawcett C, Liszkowski U (2012): Mimicry and play initiation in 18-month-old infants. Infant Behav Dev 35:689–696.
29. Carpenter M, Uebel J, Tomasello M (2013): Being mimicked increases prosocial behavior in 18-month-old infants. Child Dev 84:1511–1518.
30. Stel M, Vonk R (2010): Mimicry in social interaction: Benefits for mimickers, mimickees, and their interaction. Br J Psychol 101:311–323.
31. Oberman LM, Winkielman P, Ramachandran VS (2009): Slow echo: Facial EMG evidence for the delay of spontaneous, but not voluntary, emotional mimicry in children with autism spectrum disorders. Dev Sci 12:510–520.
32. McIntosh DN, Reichmann-Decker A, Winkielman P, Wilbarger JL (2006): When the social mirror breaks: Deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. Dev Sci 9:295–302.
33. Field T, Field T, Sanders C, Nadel J (2001): Children with autism display more social behaviors after repeated imitation sessions. Autism 5:317–323.
34. Contaldo A, Colombi C, Narzisi A, Muratori F (2016): The social effect of "being imitated" in children with autism spectrum disorder. Front Psychol 7:726.
35. Heimann M, Laberg KE, Nordøen B (2006): Imitative interaction increases social interest and elicited imitation in non-verbal children with autism. Infant Child Dev 15:297–309.
36. Folgado D, Barandas M, Matias R, Martins R, Carvalho M, Gamboa H (2018): Time alignment measurement for time series. Pattern Recognit 81:268–279.
37. Rodríguez JD, Pérez A, Lozano JA (2010): Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell 32:569–575.
38. Weinberg SL, Darlington RB (1976): Canonical analysis when number of variables is large relative to sample size. J Educ Behav Stat 1:313–332.
39. Tabachnick BG, Fidell LS (2019): Using Multivariate Statistics, 7th ed. London, United Kingdom: Pearson Education.
40. Raudys S (1997): On dimensionality, sample size, and classification error of nonparametric linear classification algorithms. IEEE Trans Pattern Anal Mach Intell 19:667–671.