# Accuracy and Reliability of At-home Quantification of Motor Impairments Using a Computer-based Pointing Task with Children with Ataxia-Telangiectasia

VINEET PANDEY, John A Paulson School of Engineering and Applied Sciences, Harvard University, USA
NERGIS C. KHAN, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, USA
ANOOPUM S. GUPTA, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, USA
KRZYSZTOF Z. GAJOS, John A Paulson School of Engineering and Applied Sciences, Harvard University, USA

Methods for obtaining accurate quantitative assessments of motor impairments are essential in accessibility research, design of adaptive ability-based assistive technologies, as well as in clinical care and medical research. Currently, such assessments are typically performed in controlled laboratory or clinical settings under professional supervision. Emerging approaches for collecting data in unsupervised settings have been shown to produce valid data when aggregated over large populations, but it is not yet established if in unsupervised settings measures of research or clinical significance can be collected accurately and reliably for individuals. We conducted a study with 13 children with ataxia-telangiectasia and 9 healthy children to analyze the validity, test-retest reliability, and acceptability of at-home use of a recent active digital phenotyping system, called Hevelius. Hevelius produces 32 measures derived from the movement trajectories of the mouse cursor, and it produces a quantitative estimate of motor impairment in the dominant arm using the dominant arm component of the Brief Ataxia Rating Scale (BARS). The severity score estimates generated by Hevelius from single at-home sessions deviated from clinician-assigned BARS scores more than the severity score estimates generated from single sessions conducted under researcher supervision. However, taking a median of as few as 2 consecutive sessions produced severity score estimates that were as accurate or better than the estimates produced from single supervised sessions. Further, aggregating as few as 2 consecutive sessions resulted in good test-retest reliability (ICC = 0.81 for A-T participants). This work demonstrated the feasibility of performing accurate and reliable quantitative assessments of individual motor impairments in the dominant arm through tasks performed at home without supervision by the researchers. Further work is needed, however, to assess how broadly these results generalize.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: active digital phenotyping, motor impairments, remote assessment, ataxia, ataxia-telangiectasia

Authors' addresses: Vineet Pandey, vineetp13@gmail.com, John A Paulson School of Engineering and Applied Sciences, Harvard University, Allston, MA, 02134, USA; Nergis C. Khan, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; Anoopum S. Gupta, agupta@mgh.harvard.edu, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; Krzysztof Z. Gajos, kgajos@seas.harvard.edu, John A Paulson School of Engineering and Applied Sciences, Harvard University, Allston, MA, 02134, USA.

1936-7228/2023/03-ART10 $15.00
https://doi.org/10.1145/3581790

**ACM Reference Format:**
Vineet Pandey, Nergis C. Khan, Anoopum S. Gupta, and Krzysztof Z. Gajos. 2023. Accuracy and Reliability of At-home Quantification of Motor Impairments Using a Computer-based Pointing Task with Children with Ataxia-Telangiectasia. *ACM Trans. Access. Comput.* 16, 1, Article 10 (March 2023), 25 pages. https://doi.org/10.1145/3581790

## 1 INTRODUCTION

Methods for obtaining accurate quantitative assessments of motor impairments are essential in accessibility research, design of adaptive ability-based assistive technologies, as well as in clinical care and medical research [15, 18, 23, 39, 68]. Currently, such assessments are typically performed in controlled laboratory or clinical settings under professional supervision. This, naturally, makes it difficult to perform such assessments at scale, longitudinally, or with populations that cannot easily access testing facilities. Further, lab-based assessments may not be representative of real-world performance [35, 54], thus limiting the value of such assessments for tailoring accessibility solutions.

To address these limitations, a number of novel approaches are being explored for motor impairment assessment in the wild [20]. One such approach is to use passive digital phenotyping. This approach relies on instrumenting a person's computer or mobile devices, or on providing the person with specialized wearable devices to collect movement data as the person naturally goes about their activities [23, 39]. Passive approaches can produce large quantities of data with minimal participant burden and, in some cases, they have produced measurements that correlate well with established clinically-relevant measures [22, 24, 29, 44]. However, attempts to collect other measures using passive approaches have been less successful [65]. Yet other passive digital phenotyping approaches have not been adapted for individuals with substantial impairments [12, 17].

Another approach for unsupervised collection of behavioral measurements (motor or cognitive) in the wild is *active* digital phenotyping, an approach that relies on participants performing specific behavioral tasks using special software or devices, but doing so on their own time, at home, and without professional supervision. Active digital phenotyping in the wild requires explicit effort on the part of the participants but it has been shown to produce results that replicate those obtained in conventional laboratory settings for both the general population [19, 31, 50, 56] as well as for individuals with unusual abilities such as children, the elderly, and people with impairments [28, 49].

These existing results demonstrate that data obtained using active digital phenotyping approaches in unsupervised settings are valid when aggregated over a large number of individuals. There is only minimal evidence so far (and only for adult participants) that these approaches can be used to obtain accurate and reliable measurements of *individual* performance [2, 9, 66]—measurements that could be used to inform individual accessibility adaptations or treatment decisions. Indeed, there are concerns that factors such as interruptions, limited motivation, or changes in environmental conditions may cause individual data collected in unsupervised settings to have lower test-retest reliability than the data collected in laboratory settings and that they are more likely to include extreme outliers [45].

To help improve our understanding of the value of unsupervised active digital phenotyping for making accurate individual assessments of motor impairments, we conducted a study to analyze the validity, test-retest reliability, and acceptability of at-home use of a recent active digital phenotyping system, called Hevelius [16]. Hevelius presents people with a simple pointing task to be performed using a computer mouse, collects complete mouse pointer movement trajectories, and produces 32 measures derived from the movement trajectories. Hevelius has been previously validated in a supervised clinical setting, where data collected under researcher supervision enabled accurate discrimination between patients and healthy individuals, and precise quantification of individual impairment in patients with ataxias and parkinsonism [16].

In this manuscript, we report on a study involving 13 children with Ataxia-telangiectasia (A-T) and 9 healthy children. The children with A-T were first assessed by a clinician and their motor impairment in the dominant arm was scored by the clinician using the Brief Ataxia Rating Scale (BARS) [60]. The children also used Hevelius once on researcher-provided equipment and under researcher supervision. Subsequently, the children used Hevelius at home—on their own computers and without researcher supervision (but typically in the presence of a care giver)—approximately once a week for up to 14 weeks.

The BARS score estimates generated by Hevelius from single at-home sessions deviated from clinician-assigned BARS scores more than the BARS score estimates generated from single sessions conducted under researcher supervision. However, taking a median of as few as 2 consecutive sessions produced severity score estimates that were as accurate or better than the estimates produced from single supervised sessions. Further, aggregating as few as 2 consecutive sessions resulted in good test-retest reliability (intraclass correlation coefficient ICC = 0.81 for A-T participants).

We also analyzed the test-retest reliability of individual measures reported by Hevelius, many of which were based on prior research on accessible computing. When the data were aggregated over 2 consecutive sessions, 6 measures showed good test-retest reliability (ICC $\geq$ 0.75) for both A-T and healthy children. These measures were the movement time, the number of pauses, duration of the longest pause, execution time (i.e., time from the first to last mouse movement event), click duration, and normalized jerk [3] (a measure of movement smoothness). Some measures, however, demonstrated poor test-retest reliability (ICC $<$ 0.50) in both A-T and healthy groups even when 4 consecutive sessions were aggregated together. These included movement offset, movement error, movement variability [41], variability in peak acceleration, maximum deviation from the task axis (i.e., maximum departure from the straight line connecting the start and end points), and the main submovement (i.e., the submovement with the highest peak speed).

Taken together, our results demonstrate that although single measurements of motor impairments collected in unsupervised settings can have higher variance than those collected in conventional laboratory settings, the simple approach of aggregating (taking a median of) a small number of consecutive unsupervised sessions can be sufficient to produce results that are as accurate or better than those obtained in supervised settings and that have high test-retest reliability. These results indicate the potential value of unsupervised active digital phenotyping for quantifying individual motor impairments in situations where multiple (longitudinal) assessments are feasible and justified. Further work is needed, however, to assess how broadly these results generalize.

## 2 BACKGROUND: ATAXIA-TELANGIECTASIA AND RELATED WORK

### 2.1 Ataxia-Telangiectasia

Ataxia-telangiectasia (A-T) is a rare, progressive, life-limiting neurological disorder. Most children with A-T do not have clear motor impairments at birth. They begin to walk at a typical age, however they do not show the same pace of gait and balance improvements that occur with typical childhood motor development [58]. Walking becomes more difficult over time and by the beginning of the second decade most children with A-T begin using a wheelchair [7]. Arm motor functions including writing, coloring, and eating become progressively more impaired during primary school years and children with A-T develop slurred speech. Involuntary movements, including chorea (jerky involuntary movements), dystonia (involuntary muscle contractions), tremor, and myoclonus (brief, involuntary muscle twitching) occur in some individuals to varying extents over the course of the disease [61]. A-T is a multisystem disorder and individuals have immunodeficiencies and increased risk for cancer [58]. The average life expectancy for individuals with A-T is approximately 25 years [8].

## 2.2 The Brief Ataxia Rating Scale (BARS)

The Brief Ataxia Rating Scale (BARS) [60] is a clinician-performed ataxia rating scale in which a clinician guides the patient through a series of motor tasks and scores performance of the task on an ordinal scale. The scale evaluates gait (natural walking and heel-to-toe tandem walking), speech (natural speech and rapid syllable production), eye movements (gaze holding, saccades, and smooth pursuit), leg movement on the heel-to-shin task, and arm movement on the finger-nose-finger task. For the finger-nose-finger task, the clinician observes the smoothness, accuracy, speed, and segmentation of the arm trajectory as the index finger goes back and forth between the clinician's finger and the patient's nose. Even though Hevelius presents people with a different task (moving the mouse pointer to click on a series of dots on a screen) than the finger-nose-finger task used for BARS, both tasks are designed to assess the underlying ability to produce and control arm movement. For that reason, we used the dominant arm component of BARS as the clinical ground truth against which we evaluated Hevelius.

## 2.3 Related Work on Digital Assessment of Motor Impairments

When it comes to characterizing people's behavioral characteristics outside of clinical or laboratory settings, we find the distinction between passive and active phenotyping [23] helpful.

In passive digital phenotyping, data are collected through specially instrumented personal devices (e.g., phones, watches, computers) or specialized wearable devices, all while people go about their natural activities [39, 55]. In some instances, passive digital phenotyping can produce informative insights. For example, keystrokes derived from typing on a laptop identified response to dopamine therapy [52]. More recently, a single wrist sensor has been shown to provide accurate, reliable, and interpretable information about the severity of motor impairments in children with A-T [44]. However, other approaches produce data that are inaccurate [65], noisy, or only allow for coarse distinctions between presence or absence of impairment [34]. Yet another approach has been to identify deliberate targeted mouse movements in the stream of a person's natural activities so that such "lab quality" movements could be used for further analyses using established techniques [12, 17]. These approaches can produce high quality assessments in some cases but, because of their reliance on models of what deliberate targeted movements should look like, they are not well suited for individuals with unusual motor abilities.

As crowdsourcing techniques became popular in research communities, numerous studies demonstrated the feasibility of collecting high quality behavioral data with remote unsupervised participants [19, 26, 31, 45, 50, 56], with several studies involving the elderly and people with impairments [28, 49]. Some of these validation studies included children (e.g., [31, 56]), however none analyzed the performance of children separately from the adults so there is some uncertainty regarding how well these approaches work for younger participants. Further, there are also results indicating that data collected on mobile devices from unsupervised participants may not be as robust as data collected on desktop computers [13]. Further still, in many of the studies, the tasks were made deliberately short (shorter than would be typical in a conventional laboratory setting) with hopes of making them acceptable to a larger numbers of participants. Thus, they traded off the quality of measurements obtained from individual participants for increased number of participants. Consequently, these studies were able to show that the aggregate results were of high quality, but they did not demonstrate whether valid and reliable assessments could be collected for individual participants. Still, these results provide a solid foundation upon which to build active digital phenotyping solutions.

In medical literature, numerous systems that have the potential to support unsupervised remote assessments of motor impairments in the hand have been proposed (see a recent systematic review [20]), but we are aware of only three that assessed the validity of the data collected in unsupervised settings for the purpose of quantifying some clinically-relevant measure of impairment [2, 9, 66]. Of those, one study with multiple sclerosis patients [9] demonstrated that an automated analysis of a patient tracing shapes with a finger on a phone screen can yield accurate predictions of how long the patient would take to complete a 9-hole peg test (a frequently used clinical assessment of motor impairments). Another study with Parkinson's patients [2] showed that an analysis of a person performing tapping and press/release tasks on a smart phone can be used to predict their UPDRS scores (UPDRS is a clinical scale for assessing the severity of symptoms of patients with Parkinson's disease). Lastly, another study with Parkinson's patients [66] tracing spirals on a screen using a stylus showed that an automated analysis of participants' drawing trajectories yielded assessments (on an ordinal scale) similar to those performed by experts. This last study was the only one of the three to also analyze the test-retest reliability of the approach.

## 3 BACKGROUND: THE HEVELIUS SYSTEM

Hevelius is a web-based tool for quantifying impairments in the dominant arm. Participants perform a set of pointing tasks with a mouse. Hevelius records detailed trajectories of the mouse pointer movements and computes 32 measures derived from prior literature on human motor performance, accessible computing, and aging. The measures are reported as age-specific z-scores by comparing participants' raw performance to the normative data collected from healthy volunteers of the same age. Because motor performance changes substantially throughout a person's lifetime (see, e.g., [14]), using age-specific z-scores makes it possible to separate the effects of a medical condition from the effects of development and aging.

Some details of the design of Hevelius have been reported in the supplementary material accompanying [16]. To make this manuscript self-contained, we describe all key design decisions of importance to this audience below. We start by describing how the normative data were collected.

### 3.1 Normative Data

To collect motor performance data from a large number of diverse healthy participants, we conducted a study with unpaid online volunteers using the LabintheWild.org platform [56]. As mentioned earlier, several validation studies have demonstrated that data collected from unpaid and unsupervised volunteers on such platforms can, in aggregate, match the quality of the data collected in conventional laboratory settings [19, 21, 31, 50, 56] (including for people with impairments [49]).

*3.1.1 Task and Procedure.* The study started with an informed consent form and brief instructions, including the request to perform the pointing tasks as quickly and as accurately as possible. Instructions were presented in English, but detailed comprehension was not essential for completing the task. During the main part of the study, participants were presented with ten blocks of eight pointing tasks. Across blocks, tasks differed in target size (10, 15, 25, 40, and 60 pixels) and distance between targets (75–400 pixels). For participants using small screens, distances between targets were scaled automatically if necessary to fit on the screen. Indices of difficulty (computed as $\log_2\left(\frac{D}{W}+1\right)$, where $D$ is the distance to the target and $W$ denotes the target diameter) varied from 2.2 to 4.8 across blocks. Half of the blocks used a reciprocal pointing task design (Figure 1 Right) and in half only one target was visible at a time: the next target appeared at fixed distance but at a random direction only after the current target was successfully acquired (Figure 1 Left). The order of the blocks was the same for all participants. The study took approximately 5 minutes.

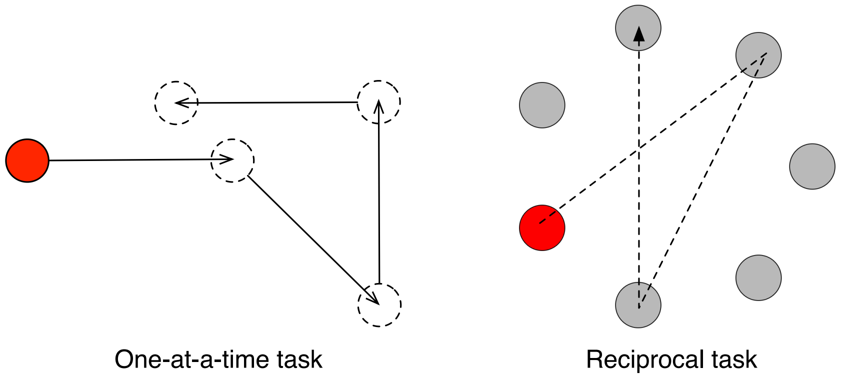


Fig. 1. Pointing tasks used to collect baseline data. Left: One-at-a-time tasks, where only one target was visible at a time and the subsequent target appeared at a fixed distance, but in a random direction. All 8 directions in 45° intervals were used once. Right: Reciprocal tasks, where all targets were visible and the location of each subsequent target was predictable.

After completing the pointing tasks, participants were asked to answer questions about their gender, input device, frequency of computer usage, and country. We also asked participants if they had any medical conditions that might affect their ability to use computers and whether they encountered any technical difficulties during the experiment. Finally, participants were asked to report their age. All questions were optional.

*3.1.2 Participants.* Approximately 540,000 people took part in the study. To develop a baseline dataset, we only included 229,017 participants who reported using a mouse, who did not report having an impairment, and who reported their age. In addition, this dataset excludes participants aged 4 or less and 86 and above, where our data became too sparse to compute meaningful baselines. We did not exclude 6% of participants who reported having encountered technical difficulties during the experiment because, after outliers were removed, we saw no significant differences on any of the measures between participants who reported having encountered technical difficulties and those who did not.

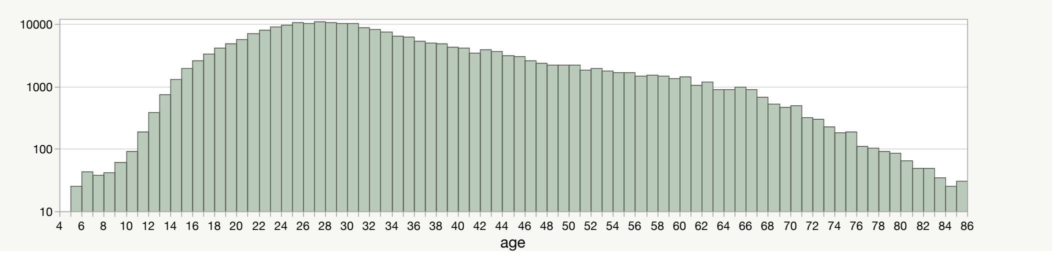


Fig. 2. Age distribution of participants who contributed to the normative data collection. Counts are shown on a logarithmic scale.

Participants included in the computation of normative baselines were between 5–85 years old ($M$ = 33.2 years, $SD$ = 12.4 years). As shown in Figure 2, young adults were well-represented in our sample (e.g., over 10,000 individuals aged 27), while many fewer children and elderly participated: the least represented were the 84-year olds (N=24) and 5-year olds (N=25). In Section 3.2.2 we

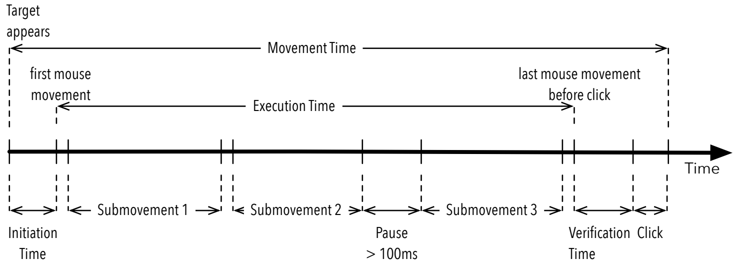


Fig. 3. Components of a movement.

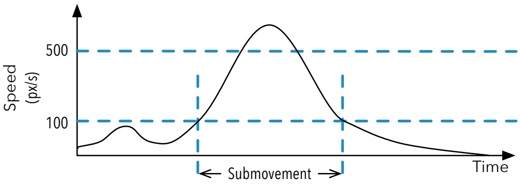


Fig. 4. Submovement identification: we used speed thresholds to identify submovements.

describe the approach we took to make the normative estimates more robust for age groups for which we had the fewest participants. 65.5% of the participants identified as male, 33.3% as female, and 1.2% chose not to disclose their gender. Participants came from 158 countries with a plurality of 43.8% coming from the U.S. and most reported using the computer for many hours on most days.

*3.1.3 Initial Processing of the Data.* We collected basic movement statistics (location of end points, timing) as well as detailed movement trajectories. Because discrete sampling of continuous mouse pointer trajectories introduces potential artifacts, we followed the general approach used by a number of other researchers (e.g., [17, 38, 53, 64, 67]) to translate, rotate, resample and smooth the data prior to computing any measures. Specifically, we first translated and rotated the movement trajectories such that each movement started at the origin and ended on the x-axis. Next, we resampled pointer position trajectories at 10 ms time intervals resulting in a 100 Hz sampling rate. We then smoothed the movement trajectories using a Kalman filter. To compute speed, we computed discrete derivative of the smoothed 2D pointer positions with respect to time and we smoothed the result using a 7Hz low-pass FIR filter (with 40dB stopband attenuation using Kaiser window). To compute acceleration and jerk, we similarly first computed discrete derivatives (of speed and acceleration, respectively) with respect to time and then applied the same low-pass filter.

Following prior work [64, 67] (and as illustrated in Figure 3), we decomposed each movement into several components: *initiation time* (from the target onset to the first mouse move event), *execution time* (from first to last mouse event), *verification time* (time spent inside the target between last mouse move event and the start of the click), and *click* (time from mouse down to mouse up event). We marked pauses whenever there was a break of 100ms or more in the raw mouse movement events. We further subdivided execution time into submovements.

Similarly to [41, 53, 67], we used speed thresholds to mark the start and end of a submovement. Specifically (as illustrated in Figure 4), a new submovement was marked when the speed crossed the 100 pixels/s threshold, but only if it subsequently reached at least 500 pixels/s. The end of the movement was marked when the speed fell again below 100 pixels/s. The *main submovement* was the submovement during which the speed reached its maximum value. Most of the time, the first submovement was the main one, but a small fraction of movements started with one or more short submovements, which were later followed by the main submovement.

*3.1.4 Raw Measures.* The 32 measures computed by Hevelius are listed in Appendix A. Most of the measures were derived from prior research in accessible computing and human motor performance [5, 6, 11, 30, 33, 41–43, 63, 64]. Hevelius was designed to be comprehensive and many measures are closely related (e.g., movement time, execution time, execution time without pauses) allowing the users of the system to choose the measures that are the most useful to their research or application.

*3.1.5 Outlier Removal.* In unsupervised online settings, extreme outlier values can be orders of magnitude different from typical values (e.g., because of a participant receiving a phone call in the middle of a trial). Because means and standard deviations are sensitive to such extreme outliers [10], we (like others [45]) used a median-based method for identifying extreme outliers. Specifically, for each raw measure we first computed the inter-quantile range between the 10th and the 90th centile (denoted as $IQR_{10-90}$) and removed all values further than $5 \times IQR_{10-90}$ from the median. This is a very conservative criterion, approximately equivalent to eliminating outliers 6.4 standard deviations from the mean for normally distributed data.

*3.1.6 Computing Age-specific Normative Baselines.* To enable computation of z-scores that are independent of a person's age and the details of the task, we performed the following computations on the normative data set:

(1) We computed per-block averages of each of the raw measures.
(2) We applied the Box-Cox transform to the per-block averages for each measure to make the distribution of the values approximately normal. Box-Cox transform [4, 59] is more general than the log-transform commonly used in HCI literature. In fact, log-transform is a special case of Box-Cox transform.
(3) For each Box-Cox transformed measure and separately for each age year, we fitted the regression of the form

$$\text{measure} = \beta_0 + \beta_1 \log_2(d) + \beta_2 \log_2(w) + \beta_3 t$$

Where $\beta_0 \ldots \beta_3$ are the parameters to be estimated, $d$ is the nominal distance between targets in the block, $w$ is the diameter of the targets, and $t$ is 0 if the task was reciprocal and 1 if it was one-at-a-time design. The effect of the task type was minimal on all measures, but we thought it was prudent to account for it.
(4) For each of the regressions (i.e., for each measure and for each age) we computed the standard deviation of the residuals.

## 3.2 Hevelius in Clinic

As previously reported [16], Hevelius was first used in a movement disorders clinic, primarily with patients with ataxias and parkinsonism. Patients used Hevelius on the same day as their scheduled visit with a neurologist, thus same-day assessments of disease severity were available.

*3.2.1 Modified tasks and procedures.* In clinic, we only used the one-at-a-time variant of the task. Unlike during the original online data collection, the in-clinic version started with 2 practice blocks to ensure that participants understood the task and that they had an opportunity to familiarize themselves with the physical set up. The main task comprised of 8 blocks of 9 trials each. As before, the first trial of each block served to position the mouse in a known position and was not included in the analyses. All participants performed the task using a standard mouse and a 17 inch display.

Target sizes varied from 16 to 90 pixels, distances between targets varied from 90 to 360 pixels, such that half of the blocks had the index of difficulty of 2, and half had the index of difficulty of 4.

*3.2.2 Computing Age-specific z-scores.* The results for each participant were reported as age-specific z-scores, separately for each measure.

The z-scores for each measure for each participant were computed by taking the difference between the observed value of a particular measure (averaged per-block and Box-Cox transformed using the same parameter that was used for the normative data) and the value estimated by the regression model for the particular task parameters and test-taker's age (see Section 3.1.6). This difference was divided by the standard deviation of the residuals of the appropriate regression model.

To account for the small number of observations in the normative data set for some ages, the overall z-scores were smoothed across neighboring ages using a locally-weighted linear regression [25] (with $\lambda = 5$). For participants who were 4 years old, this approach also allowed us to compute their z-scores by extrapolating beyond the range of ages (5 through 85) represented in the normative data set.

The z-scores were computed separately for each block of trials and later averaged across blocks.

*3.2.3 Estimates of disease severity.* Linear regression models were trained to estimate clinician-assigned disease severity scores (the dominant arm component of the Brief Ataxia Rating Scale, or BARS, for the ataxia patients; and the dominant arm component of the Unified Parkinson's Disease Rating Scale, or UPDRS for patients with parkinsonism). The models were validated using leave-one-out cross-validation. For ataxia patients, the results showed a mean average error (MAE) with respect to the clinician-assigned scores of 0.35 (on a 0–4 scale). This was a strong result given that clinicians have been estimated to have a MAE of 0.38 on the same task [40].

## 4 HEVELIUS AT HOME

For this project, we built on the version of Hevelius that had been used in a clinical setting and we adapted it for use at home without supervision by the researchers.

### 4.1 Updated pointing task

The most substantial modification we made to Hevelius to adapt it for unsupervised at-home use, was to develop a mechanism to personalize the task such that the smallest target size presented during the pointing task was appropriate to the abilities of each individual participant.

In the clinic, we had observed that some participants—particularly children with substantial impairments—were unable or unwilling to complete blocks of trials that involved very small targets. Our initial approach in clinic was to allow the supervising researcher to skip to the next block if a participant declared that they were stuck. Later, we created a capability for the supervising researcher to increase the target size and restart the block.

For this study, we used the practice tasks from the initial session—which was conducted under researcher supervision—to identify the minimum target size that each participant was capable of clicking on reliably and we used that value to personalize the tool for that person. During the rest of the supervised session and during the subsequent unsupervised at-home sessions, all the targets were set to be at least as large as the minimum target size identified during the initial session. As argued earlier, the procedure for computing z-scores in Hevelius was designed to make the results independent of the task properties (target sizes, distances between the targets, or the task type). Therefore, this adjustment—which resulted in different participants completing slightly different versions of the task—should still result in scores that are comparable across participants.

Lastly, for consistency with the earlier deployment in the clinic, we used the one-at-a-time task design by default. To keep the directions of the consecutive movements pseudo random, this task design requires a larger canvas than the reciprocal design. We instructed participants to maximize their browser windows to maximize the chances that the task would fit. However, to accommodate small screens or situations where the browser window was not maximized, Hevelius used the reciprocal as a backup if the one-at-a-time task would not fit.

### 4.2 Updated user experience

We also updated Hevelius to include a "test drive" mode to allow the caregivers to experience the entire task without the data being recorded for analysis. We also included some brief questionnaires: The first asked the caregiver to report on their perception of the participant's fatigue and cooperation levels. The second asked the participant to report on their own mood, fatigue and last night's sleep quality. Care givers were allowed to help the participants with the questionnaires but we included prominent instructions asking care givers not to help the children with the main clicking task. We further instructed them that if the task became too frustrating, they should allow the child to skip the rest of the test and that they should contact the research team to adjust the parameters of the test for their child rather than help the child with the clicking task.

### 4.3 Updated model for estimating BARS score from Hevelius measures

Lastly, taking advantage of the fact that we had additional data from in-clinic patients, we updated the regression model for estimating BARS dominant arm severity scores from measures generated by Hevelius. To create the updated model, we followed the same procedures as [16]. We extended the original data set by recruiting additional 43 ataxia patients (for a total of 138) and 7 healthy control (for a total of 36) in the same clinic in which the data for the original data were collected. As before, we used the LASSO method, which simultaneously performs feature selection and fits a linear model [62]. The new model uses values of 11 Hevelius measures (see Appendix B) compared to 5 used by the original model. Although it performed similarly to the original model when evaluated using leave-one-out cross-validation (MAE = 0.38, $r = 0.77$), we found that it was more robust when generalizing to unseen participants (who used Hevelius in the context of other studies).

## 5 STUDY

We conducted a study to assess the accuracy and test-retest reliability of the measures provided by Hevelius. Because prior work cautions that data obtained in unsupervised settings may include more numerous and more extreme outliers than data collected in conventional laboratory and clinical settings, we analyzed data obtained from single unsupervised sessions as well as data aggregated across multiple consecutive sessions. We relied on clinician-assigned BARS scores for the impairment in the dominant arm as the ground truth for evaluating the accuracy of the measurements obtained with Hevelius. We computed test-retest reliability both for the BARS estimates produced by Hevelius and for the individual measures it generates.

### 5.1 Methods

*5.1.1 Approvals.* This study was reviewed and approved by the Partners Healthcare Research Committee Institutional Review Board.

*5.1.2 Participant recruitment.* All participants were recruited in partnership with the Ataxia-telangiectasia Children's Project (A-TCP) which is a 501(c)(3) nonprofit organization that supports biomedical research projects for ataxia-telangiectasia (A-T). All recruited children with A-T were genetically confirmed to have the disorder. Children were excluded from the study if they were younger than 4 years old[1], unable to perform the computer mouse task, or had another movement disorder or condition that affected arm function or mobility. We did not impose an upper limit on the age of the participants, but because A-T is a progressive and life-limiting condition, no A-T participants older than 15 enrolled in the study. Healthy siblings of the A-T participants were recruited as healthy controls. Each participating family received one $50 American Express gift card per participating child.

#### 5.1.3 *Procedures: Supervised use.*

Participants met with researchers at an annual event organized by the A-TCP. At the event, parents provided written consent. Children 7 or older provided assent. Different assent forms were used for younger children (7 through 12) and for older children (13+). Afterwards, participants, accompanied by their caregivers, completed a session with Hevelius under researcher supervision and using a researcher-provided computer. A-T participants additionally completed a neurological exam, which was recorded on video. The video recording was later used by a clinician to quantify each participant's motor impairment in the dominant arm on the Brief Ataxia Rating Scale (BARS) [60].

While using Hevelius with a researcher, participants had the choice to request an increase in the target size in the second practice task if they felt the smallest target size (16 pixels) was too small. If such an adjustment was made for them, the selected target size was used as the minimum target size across all remaining tasks—both those completed during the rest of the supervised session and those completed later at home.

Two members of the research team were present during participants' supervised use to answer any questions. At the completion of supervised use, the research team suggested to the families that they use Hevelius at home once a week for up to 14 weeks and encouraged them to note a day and time of the week for using the tool. Researchers provided families with a USB 3 Optical Mouse [2] for at-home use. In some cases, families mentioned they were comfortable using their own mouse. Caregivers were told that they could communicate with the two members of the research team via email if they faced any issues.

#### 5.1.4 *Procedures: At-home use.*

Participants and caregivers used Hevelius without supervision on their personal computers using a mouse. The partner organization sent two emails to all participating families: 1) a reminder mail 2 weeks after their supervised use; 2) a summary of researchers' response to questions from the families. Three research team members met weekly among themselves to share weekly usage data, identify outliers, and discuss usability changes to the tool. If a family did not use the tool for two weeks, the research team updated the designated contact person at rare disease foundation whose team reached out to the caregivers (over email/phone) to understand concerns (if any).

#### 5.1.5 *Measures.*

For each participant, we collected one clinical measure: the rating (on a 0–4 scale) on the dominant arm component of the Brief Ataxia Rating Scale (BARS). This rating was assigned to A-T participants by one team member, a practicing neurologist experienced in the care of ataxia patients, based on an in-person examination and sometimes a retrospective review of the video recording of the neurological exam. All healthy participants were assigned BARS scores of 0, indicating no impairment, without clinical examination.

[1] Section 3.2.2 explains how measures were computed for 4-year-old participants even though normative data were available only for participants 5 and older.

[2] HP USB 3 Button Optical Mouse, Product number KY619AA, https://support.hp.com/sg-en/product/hp-usb-3-button-optical-mouse/3948304/document/c02576082 accessed on March 5, 2022

For each Hevelius session, we computed the individual measures and we estimated BARS scores for the dominant arm component of the scale from participants' mouse movement trajectories using the model described in Section 4.3.

For each at home session we additionally collected several subjective measures from both the caregivers and the participants. For caregivers:

- "How tired is your child right now compared to most other times?" (1=Much less tired; 5=A lot more tired)
- "How cooperative is your child right now compared to most other times?" (1=Much less cooperative; 5=Much more cooperative)

For participants:

- "What is your mood right now?" (1=🙁 (frowning face)[3], 5=😄 (smiling face))
- "How alert do you feel right now?" (1=Extremely tired; 5=Fully alert, wide awake)
- "How well did you sleep last night?" (1=😪 (sleepy face with a tear), 5=😊 (smiling face))

*5.1.6 Analyses.* We analyzed BARS rating estimates produced by Hevelius from single at-home sessions as well as estimates obtained by aggregating data from multiple consecutive sessions. We performed these aggregations by computing the median of the estimates produced by the individual sessions. We chose medians rather than means because medians are robust to extreme outliers.

We quantified the concordance between BARS rating estimates produced by Hevelius and the BARS ratings assigned by the clinician using mean absolute error (MAE). We used this measure to compare the validity of the data obtained through supervised and unsupervised use of Hevelius.

We computed the Intraclass Correlation Coefficient, or ICC (single rating, absolute-agreement, 2-way mixed-effects model) to quantify the test-retest reliability of the individual measures and the BARS rating estimates from different unsupervised Hevelius sessions. As per common heuristics [46], we interpreted the ICC using the following thresholds: 1) below 0.50: poor; 2) between 0.50 and 0.75: moderate; 3) between 0.75 and 0.90: good; 4) above 0.90: excellent.

*5.1.7 Treatment of outliers.* We included all sessions for which data was available from at least one block. We did not attempt to remove outliers. Instead, we relied on the fact that taking a median of multiple sessions makes the results robust to even extreme outliers as long as they are relatively infrequent.

## 5.2 Results

*5.2.1 Participants.* Thirty-two children, 18 with A-T and 14 healthy children, consented to participate in the study (the healthy participants were siblings of the A-T participants). Some children were consented too late to schedule a clinical exam or a supervised session. Some children ended up not using the tool at home. Ultimately, of the 18 children with A-T, 13 were included in at least one of the analyses: 10 in the analyses comparing the accuracy of the data produced during supervised and unsupervised use of Hevelius, and 11 in the test-retest reliability analyses. Table 1 shows the characteristics of the A-T participants included in analyses. Because A-T is a very rare disease, to preserve participants' anonymity, we limited the details to age and the severity of the dominant arm impairment.

Healthy participants were only included in the test-retest reliability analyses. Nine of the 14 consented healthy participants completed at least 8 at-home sessions and were thus included in the analyses. The 9 healthy participants ranged in age from 4 to 16, with the median of 11 years.

[3] The scale included only the emojis; we added textual descriptions to improve the accessibility of this manuscript.

Table 1. Summary of participants with A-T who were included in at least one of the analyses.

| Participant | Clinician-assigned BARS dominant arm score (0-4) | Age | Number of unsupervised sessions with valid data | Included in validity analysis? | Included in test-retest reliability analysis? | Number of attempted sessions | Weeks in study |
|---|---|---|---|---|---|---|---|
| P 1 | 2 | 6 | 3 | Yes | No | 4 | 4 |
| P 2 | 2 | 13 | 12 | Yes | Yes | 13 | 13 |
| P 3 | 0.5 | 7 | 13 | Yes | Yes | 13 | 14 |
| P 4 | 2 | 9 | 12 | Yes | Yes | 12 | 12 |
| P 5 | 1 | 6 | 12 | Yes | Yes | 12 | 12 |
| P 6 | 1 | 10 | 8 | Yes | Yes | 9 | 9 |
| P 7 | --- | 10 | 8 | No | Yes | 8 | 9 |
| P 8 | 2 | 15 | 12 | Yes | Yes | 14 | 14 |
| P 9 | 3 | 15 | 12 | Yes | Yes | 12 | 12 |
| P 10 | 2.5 | 12 | 9 | Yes | Yes | 9 | 10 |
| P 11 | 2.5 | 10 | 5 | Yes | No | 5 | 8 |
| P 12 | 2.5 | 10 | 10 | No | Yes | 12 | 14 |
| P 13 | --- | 4 | 12 | No | Yes | 13 | 12 |
|  | **Range: 0.5 - 3** | **Range: 4 - 15** | **Range: 3 - 13** | **N = 10** | **N = 11** | **Range: 4 - 14** | **Range: 4 - 14** |
|  | **Median: 2** | **Median: 10** | **Median: 12** |  |  | **Median: 12** | **Median: 12** |

The minimum target sizes set for A-T participants ranged from 16 pixels (the smallest possible) to 83 pixels, with a median of 30 pixels. Most sessions (81%) completed by A-T participants involved the one-at-a-time task design. No participant had more than 33% of sessions that utilized the reciprocal design indicating that the one-at-a-time design could fit on all participants' screens but the guidance we provided for making the browser window full screen was perhaps not adequate.

*5.2.2 Preliminary analyses: Sensitivity to task parameters.* As described in Sections 3.1.6 and 3.2.2, the z-scores for the individual measures and, consequently, the BARS estimates produced by Hevelius are designed to be independent of the task properties (i.e., the target size, the distance between targets, and whether the task uses the reciprocal or the one-at-a-time design). Given that different participants had different minimum target sizes and that there was some variability in the task types, before proceeding with the main analyses, we tested whether the assumption that the BARS estimates were independent of task properties held for the population that participated in this study.

Specifically, we used mixed effects regression models to analyze the association between the session task properties (task type and mean target size) and the estimated BARS scores. Because there was no variability either between or within subjects in distances between targets, we did not include distance to target as a factor. We used the clinician-assigned BARS scores to control for disease severity and we modelled individual participants as random effects to surface within-subjects effects.

As shown in Table 2, there was no statistically significant effect of either task type or the target size on the BARS score estimates. This held when all participants were modeled together (A-T and Healthy Model 1) or when A-T and Healthy participants were modeled separately. Of course, lack of a statistically significant effect is not sufficient evidence to conclude that the effect does not exist. Thus, we also explicitly compared a model that contained task properties (Model 1) to a model that did not (Model 0) and we quantified the difference in the amount of variance explained by the two models (as represented by the $R^2$). The results show that adding the task properties only minimally increased the amount of variance explained ($\Delta R^2$(marginal) $< .01$) and the difference between the models was not significant ($\chi^2(2, N = 206) = 3.27, n.s.$). The very small $\Delta R^2$ gives us

Table 2. Mixed-effects models capturing associations between task properties and the estimated BARS scores (controlling for clinician-assigned BARS scores). Participants are modeled as random effects. $R^2$ (marginal) captures the variance explained by fixed effects only. $R^2$ (conditional) captures the variance explained by fixed and random effects together.

| | A-T and Healthy | | A-T only | Healthy only |
|---|---|---|---|---|
| | Model 0 | Model 1 | | |
| (Intercept) | $0.45\ (0.10)^{***}$ | $0.66\ (0.52)$ | $1.37\ (0.74)$ | $-0.01\ (0.65)$ |
| Clinician-assigned BARS score | $0.73\ (0.07)^{***}$ | $0.74\ (0.07)^{***}$ | $0.45\ (0.13)^{***}$ | |
| Task type: one-at-a-time | | $0.11\ (0.07)$ | $0.14\ (0.09)$ | $0.04\ (0.07)$ |
| $log_2$(mean target size) | | $-0.06\ (0.09)$ | $-0.07\ (0.12)$ | $0.06\ (0.12)$ |
| $R^2$ (marginal) | .77 | .77 | .34 | .01 |
| $R^2$ (conditional) | .88 | .88 | .58 | .54 |
| Num. observations (sessions) | 206 | 206 | 118 | 88 |
| Num. groups (participants) | 20 | 20 | 11 | 9 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

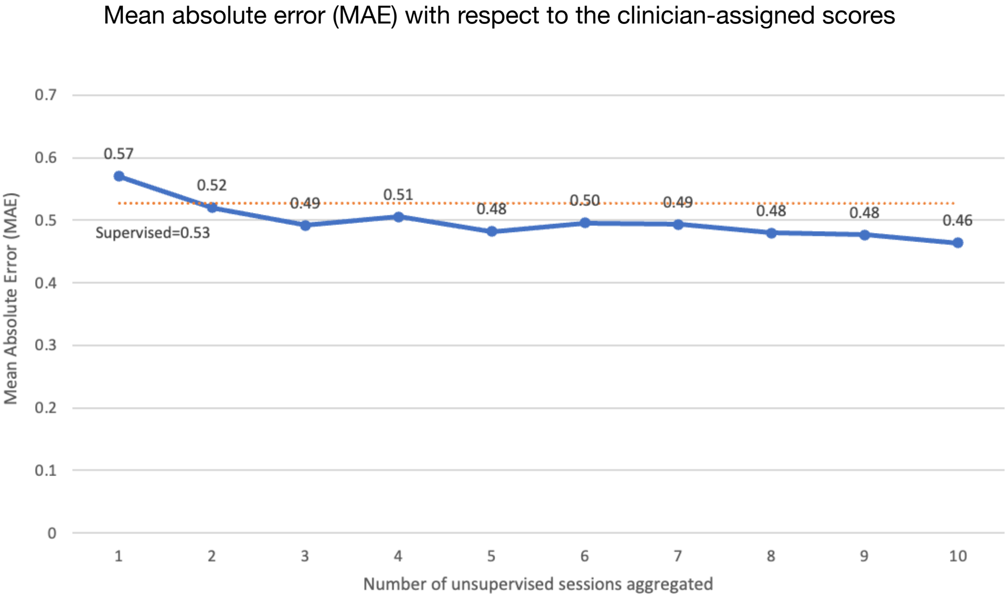


Fig. 5. Concordance with clinician-assigned scores measured using the mean absolute error (MAE).

confidence that even if the task properties impacted the BARS score estimates, the impact was of little practical significance.

*5.2.3 Comparison of the accuracy of measurements obtained from supervised and unsupervised sessions.* As a reminder, each participant completed a single session under the direct supervision

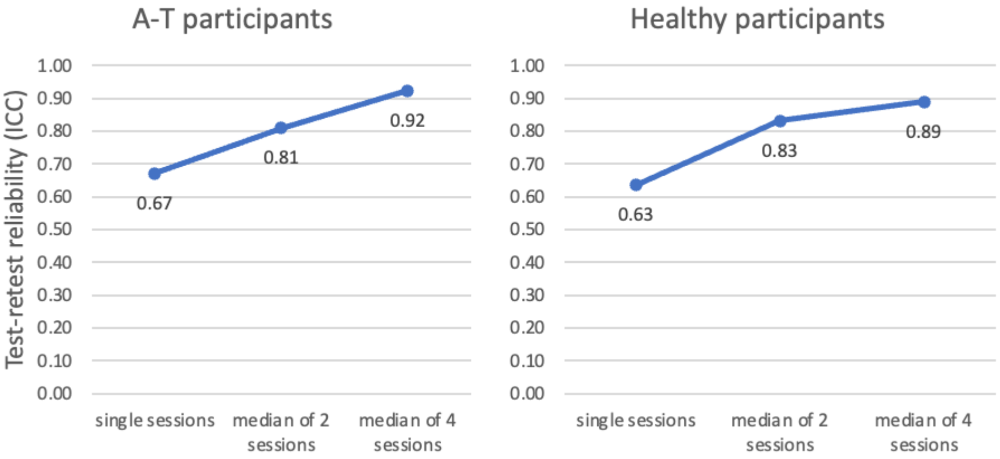


Fig. 6. Test-retest reliability. Left: for A-T participants (N=12). Right: for healthy participants (N=9).

of a researcher. In this section we compare the results from aggregating one or more consecutive unsupervised sessions (always starting with the first one) to the results obtained from each participant's single supervised session.

As expected, the mean absolute error (MAE) computed with respect to the clinician-assigned scores was higher for a single unsupervised session than for the single supervised session (MAE=.57 for a single unsupervised session; MAE = .53 for the supervised session). However, as shown in Figure 5, after aggregating just 2 consecutive unsupervised sessions, the MAE from unsupervised sessions was consistently lower than the MAE from the 1 supervised session.

These results are also promising in absolute terms: the MAE for 2 or more aggregated sessions was between .46 and .52. In comparison, expert clinicians have been estimated to have a MAE of .38 on the same task [40].

*5.2.4 Test-retest reliability.* We computed the test-retest reliability using ICC for the first 8 sessions in three ways: for single sessions (8 measurements per participant), using medians of pairs of consecutive sessions (resulting in 4 measurements per participant), and using medians of 4 consecutive sessions (resulting in 2 measurements per participant).

As shown in Figure 6, for both A-T and healthy participants, the test-retest reliability was moderate when measurements represented single sessions. However, for both groups the ICC was good (i.e., $\geq .75$) once a median of 2 consecutive sessions were used. When a median of 4 sessions was used, the results were excellent (ICC $\geq .90$) for the A-T participants and nearly excellent for the healthy participants.

Table 3 shows the test-retest reliability for all 32 measures generated by Hevelius. We show the results separately for A-T and healthy participants and for different levels of aggregation (single sessions, 2 consecutive sessions, 4 consecutive sessions). When 2 consecutive sessions were aggregated, 6 measures showed good test-retest reliability (ICC $\geq .75$) for both A-T and healthy children. These measures were the movement time, the number of pauses, duration of the longest pause, execution time (i.e., time from the first to last mouse movement event), click duration, and normalized jerk [3] (a measure of movement smoothness). Some measures, however, demonstrated poor test-retest reliability (ICC $< .50$) in both A-T and healthy groups even when 4 consecutive sessions were aggregated together. These were: movement offset, movement error,

Table 3. Test-retest reliability for all measures computed by Hevelius (using ICC). E = Excellent test-retest reliability (ICC ≥ .90), G = Good (.90 > ICC ≥ .75), M = Moderate (.75 > ICC ≥ .50), P = Poor (.50 > ICC). Analyses are provided separately for A-T participants and healthy participants, and separately for three aggregation levels: single sessions, medians of two consecutive sessions, and medians of 4 consecutive sessions.

| | A-T participants | | | | | | Healthy participants | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Measure** | **single sessions** | | **2 sessions** | | **4 sessions** | | **single sessions** | | **2 sessions** | | **4 sessions** | |
| Movement time | 0.94 | E | 0.97 | E | 0.98 | E | 0.83 | G | 0.90 | E | 0.92 | E |
| Number of pauses | 0.87 | G | 0.97 | E | 0.97 | E | 0.81 | G | 0.86 | G | 0.91 | E |
| Duration of the longest pause | 0.86 | G | 0.97 | E | 0.97 | E | 0.81 | G | 0.86 | G | 0.90 | E |
| Execution time | 0.77 | G | 0.88 | G | 0.91 | E | 0.79 | G | 0.90 | E | 0.94 | E |
| Click duration | 0.73 | M | 0.82 | G | 0.88 | G | 0.80 | G | 0.91 | E | 0.94 | E |
| Normalized jerk | 0.65 | M | 0.79 | G | 0.84 | G | 0.82 | G | 0.91 | E | 0.94 | E |
| Execution time (w/o pauses) variability | 0.32 | P | 0.49 | P | 0.84 | G | 0.45 | P | 0.55 | M | 0.82 | G |
| Execution time variability | 0.31 | P | 0.44 | P | 0.82 | G | 0.60 | M | 0.74 | M | 0.88 | G |
| Verification time variability | 0.39 | P | 0.52 | M | 0.80 | G | 0.75 | G | 0.85 | G | 0.90 | E |
| Verification time | 0.55 | M | 0.60 | M | 0.79 | G | 0.85 | G | 0.90 | E | 0.90 | E |
| Execution time (w/o pauses) | 0.47 | P | 0.62 | M | 0.78 | G | 0.71 | M | 0.84 | G | 0.93 | E |
| Click slip | 0.49 | P | 0.58 | M | 0.76 | G | 0.52 | M | 0.60 | M | 0.75 | G |
| Normalized jerk (w/o pauses) | 0.42 | P | 0.58 | M | 0.72 | M | 0.69 | M | 0.80 | G | 0.90 | E |
| Target reentries | 0.34 | P | 0.48 | P | 0.69 | M | 0.61 | M | 0.75 | G | 0.65 | M |
| Movement direction changes | 0.52 | M | 0.68 | M | 0.66 | M | 0.14 | P | 0.23 | P | 0.30 | P |
| Distance from target center at the end of the main submovement | 0.36 | P | 0.49 | P | 0.64 | M | 0.58 | M | 0.69 | M | 0.67 | M |
| Orthogonal direction changes | 0.39 | P | 0.51 | M | 0.56 | M | 0.19 | P | 0.38 | P | 0.53 | M |
| Peak speed | 0.32 | P | 0.45 | P | 0.54 | M | 0.57 | M | 0.72 | M | 0.85 | G |
| Peak acceleration | 0.33 | P | 0.45 | P | 0.51 | M | 0.55 | M | 0.71 | M | 0.87 | G |
| Click duration variability | 0.36 | P | 0.42 | P | 0.48 | P | 0.20 | P | 0.39 | P | 0.61 | M |
| Fraction of the main submovement spent accelerating | 0.26 | P | 0.36 | P | 0.43 | P | 0.24 | P | 0.45 | P | 0.67 | M |
| Peak speed variability | 0.31 | P | 0.38 | P | 0.40 | P | 0.23 | P | 0.45 | P | 0.73 | M |
| Main submovement | 0.27 | P | 0.40 | P | 0.37 | P | -0.03 | P | -0.08 | P | -0.31 | P |
| Movement time variability | 0.10 | P | 0.17 | P | 0.36 | P | 0.62 | M | 0.79 | G | 0.81 | G |
| Task axis crossings | 0.35 | P | 0.40 | P | 0.35 | P | 0.35 | P | 0.51 | M | 0.52 | M |
| Peak acceleration variability | 0.09 | P | 0.10 | P | 0.31 | P | 0.17 | P | 0.38 | P | 0.47 | P |
| Fraction of the distance to the target center covered in main submovement | 0.00 | P | 0.00 | P | 0.23 | P | 0.25 | P | 0.43 | P | 0.64 | M |
| Max deviation from task axis | 0.15 | P | 0.24 | P | 0.15 | P | 0.21 | P | 0.30 | P | 0.24 | P |
| Movement offset | 0.08 | P | 0.09 | P | 0.12 | P | 0.20 | P | 0.33 | P | -0.03 | P |
| Movement error | 0.18 | P | 0.30 | P | 0.11 | P | 0.21 | P | 0.33 | P | 0.27 | P |
| Movement variability | 0.16 | P | 0.24 | P | 0.07 | P | 0.24 | P | 0.35 | P | 0.36 | P |
| Noise to force ratio | 0.08 | P | 0.06 | P | -0.04 | P | 0.34 | P | 0.57 | M | 0.86 | G |

movement variability [41], variability in peak acceleration, maximum deviation from the task axis (i.e., maximum departure from the straight line connecting the start and end points), and the main submovement (i.e., the submovement with the highest peak speed).

*5.2.5 Do Caregiver and Participant Self-reports Explain Session-to-session Variability?* One possible explanation for between-session variability in performance might be the day-to-day changes in a participant's mood and fatigue. Thus, we conducted a regression analysis to test for the association between participant state (as reported by the caregiver and the participant) and the Hevelius-estimated BARS scores. We used mixed effects models with individual participants modeled as random effects. One model included only a covariate (the clinician-assigned BARS score). The other also included the values of all subjective measures reported by both the caregivers and the

Table 4. Mixed effects models to measure the association between subjective measures of A-T participant mood and fatigue, and the per-session Hevelius-estimated BARS scores.

| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | 1.07 (0.27)$^{***}$ | 0.96 (0.57) |
| Clinician-assigned BARS score | 0.45 (0.13)$^{***}$ | 0.45 (0.13)$^{***}$ |
| Caregiver report: participant tired | | 0.08 (0.06) |
| Caregiver report: participant cooperative | | 0.03 (0.07) |
| Participant self-report: mood | | −0.01 (0.07) |
| Participant self-report: alert | | −0.05 (0.07) |
| Participant self-report: sleep quality | | 0.00 (0.06) |
| $R^2$ (marginal) | .33 | .33 |
| $R^2$ (conditional) | .57 | .58 |
| Num. observations (sessions) | 118 | 118 |
| Num. groups (participants) | 11 | 11 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

participants. The results are summarized in Table 4. Analyzing data from 118 sessions performed by 11 A-T participants for whom we had the clinician-assigned BARS scores, we observed no significant difference in goodness of fit between the two models ($\Delta R^2 < 0.02$, $\chi^2(5, N = 118) = 4.25$, n.s.). This indicates that the subjective reports by caregivers and participants did not substantially explain the session-to-session differences in measurements.

## 5.3 Acceptability of at-home assessments by participants with A-T

*5.3.1 Time burden.* Ninety-five percent of the sessions completed by the A-T participants took between 2:38 and 28:04 minutes (counting from the start of the practice task to the end of the last block), with median session taking 11:02 minutes.

*5.3.2 Challenges in using Hevelius.* Several caregivers and participants noted that they would have preferred a shorter task. Some indicated that the task could be frustratingly difficult:

> "[Participant] move (sic) the mouse back and forth across the screen in frustration, as I'm sure you'll see in the results"

One caregiver also noted that the mouse was difficult for a participant to use:

> "the mouse is difficult to use, as his fingers keep hitting the rolling piece in the middle which causes google chrome to ask if we want to close out the program. The combination becomes frustrating for him."

Three families also mentioned health and lifestyle challenges in using the tool regularly. Lifestyle concerns included travel and sports tournaments. Some caregivers provided suggestions to improve the assessment such as including audio feedback and gamifying the task.

*5.3.3 Participants developed strategies to perform the task.* For 22 sessions, caregivers reported that A-T participants altered their sitting posture while performing the pointing task. One common strategy that emerged in the data was that participants used their non-dominant hand (hand not used for the pointing task) to stabilize themselves. Specifically, caregivers reported that participants used their non-dominant hand to brace themselves on the on chair/bench they were sitting on; to steady the wrist of the main hand; or to hold the table on which the laptop with the task was used.

Another strategy was to lean in closer to the laptop: 3 caregivers reported that participants leaned forward to the screen (presumably) to see more clearly, especially for smaller targets.

## 6 DISCUSSION, FUTURE DIRECTIONS, AND CONCLUSION

We set out to empirically evaluate the validity, test-retest reliability, and acceptability of at-home use of Hevelius, a system for quantifying motor impairments in the dominant arm [16]. Hevelius presents people with a simple pointing task, collects complete movement trajectories, and produces 32 measures derived from the movement trajectories. Based on a previous in-clinic deployment of Hevelius, a regression model has been created and validated [16] for estimating the dominant arm component of the Brief Ataxia Rating Scale (BARS), a clinical rating scale used for assessing the severity of ataxia symptoms. We conducted our evaluation with 13 children with Ataxia-telangiectasia (A-T) and 9 healthy children, who used Hevelius once under a researcher supervision and then used it at home, roughly once a week for up to 14 weeks, with the assistance of an adult caregiver but without any direct supervision from the research team.

As expected, data from a single unsupervised session matched the clinician-assigned scores less accurately than the data obtained during a single session supervised by a researcher. However, aggregating data from just two consecutive unsupervised sessions was sufficient to make the BARS estimates as accurate as those obtained in a supervised setting. Similarly, aggregating data from just two consecutive sessions was sufficient to achieve good test-retest reliability for the estimated BARS scores and for 6 individual measures produced by Hevelius. Increasing the number of consecutive unsupervised sessions that were aggregated together further improved the accuracy of the score estimates and the test-retest reliability of both BARS score estimates and of the individual measures. Given the possible occurrence of extreme outliers in data collected in unsupervised settings (e.g., because a person gets interrupted by an external event), taking a median of at least three consecutive sessions is advisable and, in the context of our study, sufficient.

We examined the possibility that the session-to-session variability in each participant's performance could be explained at least in part by changes in participants' mood and energy levels. However, neither caregiver reports of participant state nor the participant self-reports (nor the combination of the two) were significantly associated with within-participant session-to-session differences in BARS score estimates. In future work, we will collect more qualitative data to try to identify possible causes of session-to-session variability.

Our data also indicate that some of our participants did not enjoy completing the tasks. Making the task shorter could make it a little more acceptable to the participants but it would also reduce the quality of the measurements. Instead, we believe that the success of future deployments will rely on better engaging participants' motivation. Prior work on engaging healthy adults with behavioral research demonstrated the effectiveness of curiosity [48] as well as social comparison [32] in motivating participation. These mechanisms have been operationalized by creating opportunities for participants to view their own results and to compare themselves to others [56]. We chose not to employ those mechanisms because A-T is a progressive life-limiting disease and some families of A-T children do not wish to always think about changes in their child's health status [37]. Per our participants' suggestions, in our future work, we are likely to resort to entertainment as a mechanism to encourage participation. Some researchers have explored gamifying the primary assessment or rehabilitation tasks for children [36] but concerns remain that it is challenging to make such tasks both valid and entertaining at the same time. For that reason, we are unlikely to redesign the core Hevelius task itself and instead will explore adding small elements of entertainment such as brief animations with popular cartoon characters between study blocks, or fun game-like activities at the end of each session.

Although we did not specifically ask about it, the weekly assessments likely increased the work load for many caregivers who already devote a lot of effort to their care giving responsibilities [1]. To reduce the number and frequency of explicit measurements, one future direction would be to develop methodologies that combine active and passive phenotyping. As a reminder, Hevelius exemplifies active digital phenotyping as it requires participants to perform carefully specified tasks while the measurements are being collected. Passive phenotyping techniques use mobile phones or specialized wearable devices to unobtrusively collect data while a person goes about their natural activities. A solution that combined both approaches could leverage active phenotyping for infrequent but accurate measurements that could be used to calibrate and interpret more frequent data from passive measurements.

An important aspect of Hevelius is that it reports measurements as z-scores that are independent of the details of task properties and that can be age-specific. This was enabled by a large normative data set. This aspect of Hevelius makes it possible for the detailed task properties (i.e., target sizes, distances) to be adjusted within a small range to fit the available hardware and the abilities of the participants. When measuring impact of the disease—as was the case in this study—the age-specific scores also enable the separation of the effects of the disease from the effects of development and aging. If the purpose is to make accessibility adjustments that take into consideration *both* age and medical conditions, the z-scores for all participants can be computed against a common baseline instead.

The results of our study also revealed that some measures of motor performance that are used in our community may have much higher test-retest reliability than others for the specific populations represented in our study.

A key limitation of our work is that it was conducted in the context of a single disease and a narrow age group. From that perspective, our results should be interpreted as initial evidence to be extended with other medical conditions and populations. We speculate, however, that young children may be a particularly challenging population because of the difficulties they experience in persevering on boring tasks with obvious benefits. Therefore, future studies with adult participants are likely to produce stronger results particularly with respect to the test-retest reliability.

The above limitations notwithstanding, we consider it a strength of this work that it engaged with a rare disease and a pediatric population, both of which have relatively little representation in the literature. We also note that techniques for performing accurate quantitative assessments of motor behavior have the potential to be particularly valuable for supporting remote care, research, and clinical trials for patients with rare neurological disorders given that most such patients live far away from specialists knowledgeable about their disease [37].

Participants in our study used their own computers, likely of different kinds and potentially with different mouse gain settings, pixel ratios and other parameter differences. Our analysis of the normative data indicated that the measures were not substantially sensitive to such differences as long as a computer mouse was used as input.

Prior research has noted the difficulty of finding sufficient numbers of research participants with rare disorders [47]. This study was possible because of the support from the Ataxia-Telangiectasia Children's Project (A-TCP), a rare disease foundation. Having worked with the A-T patient community for many years, A-TCP supported this project in multiple ways. First, the foundation publicized the study to the member families and helped find participants. Second, the supervised use of our tool happened at an annual gathering of families of children with A-T organized by A-TCP. Meeting multiple knowledgeable and interested families in one place would have been challenging otherwise. Additionally, such initial face-to-face interactions can also improve trust in electronic contexts [57]. Third, the foundation staff reminded participating families when they did not use Hevelius (after receiving usage updates from the research team).

To conclude, this work demonstrated the feasibility of performing accurate and reliable quantitative assessments of motor impairments in the dominant arm through tasks performed at home without supervision by the researchers. Further work is needed, however, to assess how broadly these results generalize.

## ACKNOWLEDGMENTS

We thank Jennifer Thornton and Sara Reiling from A-TCP for helping the research team recruit families for the study and communicating with them. We thank Aleksandra Koralczyk for conducting an interview with some families. Vineet thanks Herman Saksono for his inputs framing this work for an HCI audience. This research was supported in part by the Sony Group Corporation (award A47628), the Ataxia-Telangiectasia Children's Project, Biogen, and NIH R01 NS117826.

The authors declare that they have no relevant conflicts of interest or financial disclosures.

## REFERENCES

[1] Ronald D Adelman, Lyubov L Tmanova, Diana Delgado, Sarah Dion, and Mark S Lachs. 2014. Caregiver burden: a clinical review. *Jama* 311, 10 (2014), 1052–1060.

[2] Siddharth Arora, Vinayak Venkataraman, Andong Zhan, S Donohue, Kevin M Biglan, E Ray Dorsey, and Max A Little. 2015. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. *Parkinsonism & related disorders* 21, 6 (2015), 650–653.

[3] Sivakumar Balasubramanian, Alejandro Melendez-Calderon, and Etienne Burdet. 2012. A robust and sensitive metric for quantifying movement smoothness. *IEEE Transactions on Biomedical Engineering* 59, 8 (2012), 2126–2136. https://doi.org/10.1109/TBME.2011.2179545

[4] George EP Box and David R Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 26, 2 (1964), 211–252.

[5] Yuenkeen Cheong, Randa L Shehab, and Chen Ling. 2013. Effects of age and psychomotor ability on kinematics of mouse-mediated aiming movement. *Ergonomics* 56, 6 (2013), 1006–1020.

[6] J D Cooke, S H Brown, and D A Cunningham. 1989. Kinematics of arm movements in elderly humans. *Neurobiology of aging* 10, 2 (March 1989), 159–165.

[7] Thomas O Crawford. 1998. Ataxia telangiectasia. In *Seminars in pediatric neurology*, Vol. 5. Elsevier, 287–294.

[8] Thomas O Crawford, RL Skolasky, R Fernandez, KJ Rosquist, and HM Lederman. 2006. Survival probability in ataxia telangiectasia. *Archives of disease in childhood* 91, 7 (2006), 610–611.

[9] AP Creagh, C Simillion, A Scotland, F Lipsmeier, C Bernasconi, S Belachew, J Van Beek, M Baker, C Gossens, M Lindemann, et al. 2020. Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the Draw a Shape Test. *Physiological Measurement* 41, 5 (2020), 054002. https://doi.org/10.1088/1361-6579/ab8771

[10] Jay L Devore. 2011. *Probability and Statistics for Engineering and the Sciences.* Cengage learning.

[11] Afke Donker and Pieter Reitsma. 2007. Aiming and clicking in young children's use of the computer mouse. *Computers in human behavior* 23, 6 (Nov. 2007), 2863–2874.

[12] Abigail Evans and Jacob Wobbrock. 2012. Taming Wild Behavior: The Input Observer for Obtaining Text Entry and Mouse Pointing Measures from Everyday Computer Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12).* ACM, New York, NY, USA, 1947–1956. https://doi.org/10.1145/2207676.2208338

[13] Leah Findlater, Joan Zhang, Jon E. Froehlich, and Karyn Moffatt. 2017. Differences in Crowdsourced vs. Lab-Based Mobile and Desktop Input Performance Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17).* Association for Computing Machinery, New York, NY, USA, 6813–6824. https://doi.org/10.1145/3025453.3025820

[14] Leah Findlater and Lotus Zhang. 2020. Input Accessibility: A Large Dataset and Summary Analysis of Age, Motor Ability and Input Performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20).* Association for Computing Machinery, New York, NY, USA, Article 17, 6 pages. https://doi.org/10.1145/3373625.3417031

[15] Krzysztof Z. Gajos, Amy Hurst, and Leah Findlater. 2012. Personalized dynamic accessibility. *interactions* 19, 2 (March 2012), 69–73. https://doi.org/10.1145/2090150.2090167

[16] Krzysztof Z. Gajos, Katharina Reinecke, Mary Donovan, Christopher D. Stephen, Albert Y. Hung, Jeremy D. Schmahmann, and Anoopum S. Gupta. 2020. Computer Mouse Use Captures Ataxia and Parkinsonism, Enabling Accurate Measurement and Detection. *Movement Disorders* 35 (February 2020), 354–358. Issue 2. https://doi.org/10.1002/mds.27915

[17] Krzysztof Z. Gajos, Katharina Reinecke, and Charles Herrmann. 2012. Accurate Measurements of Pointing Performance from in Situ Observations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 3157–3166. https://doi.org/10.1145/2207676.2208733

[18] Krzysztof Z. Gajos, Daniel S. Weld, and Jacob O. Wobbrock. 2010. Automatically generating personalized user interfaces with Supple. *Artificial Intelligence* 174 (2010), 910–950. Issue 12–13. https://doi.org/10.1016/j.artint.2010.05.005

[19] Laura Germine, Ken Nakayama, Bradley C Duchaine, Christopher F Chabris, Garga Chatterjee, and Jeremy B Wilmer. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic bulletin & review* 19, 5 (2012), 847–857.

[20] Arpita Gopal, Wan-Yu Hsu, Diane D Allen, and Riley Bove. 2022. Remote Assessments of Hand Function in Neurological Disorders: Systematic Review. *JMIR Rehabil Assist Technol* 9, 1 (9 Mar 2022), e33157. https://doi.org/10.2196/33157

[21] Samuel D Gosling, Simine Vazire, Sanjay Srivastava, and Oliver P John. 2004. Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist* 59, 2 (2004), 93–104. https://doi.org/10.1037/0003-066X.59.2.93

[22] Robert I Griffiths, Katya Kotschet, Sian Arfon, Zheng Ming Xu, William Johnson, John Drago, Andrew Evans, Peter Kempster, Sanjay Raghav, and Malcolm K Horne. 2012. Automated assessment of bradykinesia and dyskinesia in Parkinson's disease. *Journal of Parkinson's disease* 2, 1 (2012), 47–55.

[23] Anoopum S. Gupta. 2022. Digital Phenotyping in Clinical Neurology. *Seminars in Neurology* 42 (2022), 48–59. https://doi.org/10.1055/s-0041-1741495

[24] Anoopum S Gupta, Anna C Luddy, Nergis C Khan, Sara Reiling, and Jennifer Karlin Thornton. 2022. Real-life Wrist Movement Patterns Capture Motor Impairment in Individuals with Ataxia-Telangiectasia. *The Cerebellum* (2022), 1–11. https://doi.org/10.1007/s12311-022-01385-5

[25] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

[26] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems* (Atlanta, Georgia, USA) *(CHI '10)*. ACM, New York, NY, USA, 203–212. https://doi.org/10.1145/1753326.1753357

[27] Neville Hogan and Dagmar Sternad. 2009. Sensitivity of smoothness measures to movement duration, amplitude, and arrests. *Journal of motor behavior* 41, 6 (Nov. 2009), 529–534.

[28] Andrew Hooyman, Joshua S Talboom, Matthew D DeBoth, Lee Ryan, Matthew J Huentelman, and Sydney Y Schaefer. 2021. Remote, unsupervised functional motor task evaluation in older adults across the United States using the MindCrowd electronic cohort. *Developmental Neuropsychology* 46, 6 (2021), 435–446.

[29] Malcolm K Horne, Sarah McGregor, and Filip Bergquist. 2015. An objective fluctuation score for Parkinson's disease. *PloS one* 10, 4 (2015), e0124522.

[30] Juan Pablo Hourcade, Benjamin B. Bederson, Allison Druin, and François Guimbretière. 2004. Differences in pointing task performance between preschool children and adults using mice. *ACM Trans. Comput.-Hum. Interact.* 11, 4 (2004), 357–386. https://doi.org/10.1145/1035575.1035577

[31] Bernd Huber and Krzysztof Z Gajos. 2020. Conducting online virtual environment experiments with uncompensated, unsupervised samples. *Plos one* 15, 1 (2020), e0227629.

[32] Bernd Huber, Katharina Reinecke, and Krzysztof Z. Gajos. 2017. The Effect of Performance Feedback on Social Media Sharing at Volunteer-Based Online Experiment Platforms. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 1882–1886. https://doi.org/10.1145/3025453.3025553

[33] Amy Hurst, Scott E. Hudson, Jennifer Mankoff, and Shari Trewin. 2008. Automatically Detecting Pointing Performance. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (Gran Canaria, Spain) *(IUI '08)*. Association for Computing Machinery, New York, NY, USA, 11–19. https://doi.org/10.1145/1378773.1378776

[34] Amy Hurst, Scott E. Hudson, Jennifer Mankoff, and Shari Trewin. 2013. Distinguishing Users By Pointing Performance in Laboratory and Real-World Tasks. *ACM Trans. Access. Comput.* 5, 2, Article 5 (Oct. 2013), 27 pages. https://doi.org/10.1145/2517039

[35] Amy Hurst, Jennifer Mankoff, and Scott E. Hudson. 2008. Understanding pointing problems in real world computing environments. In *Assets '08: Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility* (Halifax, Nova Scotia, Canada). ACM, New York, NY, USA, 43–50. https://doi.org/10.1145/1414471.1414481

[36] Marco Iosa, Cristiano Maria Verrelli, Amalia Egle Gentile, Martino Ruggieri, and Agata Polizzi. 2022. Gaming technology for pediatric neurorehabilitation: A systematic review. *Frontiers in Pediatrics* 10 (2022).

[37] Maia Jacobs, Galina Gheihman, Krzysztof Z. Gajos, and Anoopum S. Gupta. 2019. "I Think We Know More than Our Doctors": How Primary Caregivers Manage Care Teams with Limited Disease-Related Expertise. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 159 (nov 2019), 22 pages. https://doi.org/10.1145/3359261

[38] R.J. Jagacinski, D.W. Repperger, M.S. Moran, S.L. Ward, and B. Glass. 1980. Fitts' law and the microstructure of rapid discrete movements. *Journal of Experimental Psychology: Human Perception and Performance* 6, 2 (1980), 309–320.

[39] Sachin H Jain, Brian W Powers, Jared B Hawkins, and John S Brownstein. 2015. The digital phenotype. *Nature biotechnology* 33, 5 (2015), 462–463.

[40] Ronnachai Jaroensri, Amy Zhao, Guha Balakrishnan, Derek Lo, Jeremy D Schmahmann, Frédo Durand, and John Guttag. 2017. A Video-Based Method for Automatically Rating Ataxia. In *Proceedings of the 2nd Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 68)*, Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (Eds.). PMLR, 204–216. https://proceedings.mlr.press/v68/jaroensri17a.html

[41] Simeon Keates, Faustina Hwang, Patrick Langdon, P. John Clarkson, and Peter Robinson. 2002. Cursor measures for motion-impaired computer users. In *Assets '02: Proceedings of the fifth international ACM conference on Assistive technologies* (Edinburgh, Scotland). ACM, New York, NY, USA, 135–142. https://doi.org/10.1145/638249.638274

[42] Simeon Keates and Shari Trewin. 2005. Effect of age and Parkinson's disease on cursor positioning using a mouse. In *Assets '05: Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*. ACM Press, New York, NY, USA, 68–75. https://doi.org/10.1145/1090785.1090800

[43] Caroline J Ketcham, Rachael D Seidler, Arend W A Van Gemmert, and George E Stelmach. 2002. Age-related kinematic differences as influenced by task difficulty, target size, and movement amplitude. *J Gerontol B Psychol Sci Soc Sci* 57, 1 (Jan. 2002), P54–64.

[44] Nergis C Khan, Vineet Pandey, Krzysztof Z Gajos, and Anoopum S Gupta. 2022. Free-Living Motor Activity Monitoring in Ataxia-Telangiectasia. *The Cerebellum* 21, 3 (2022), 368–379. https://doi.org/10.1007/s12311-021-01306-y

[45] Steven Komarov, Katharina Reinecke, and Krzysztof Z. Gajos. 2013. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 207–216. https://doi.org/10.1145/2470654.2470684

[46] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15, 2 (2016), 155–163.

[47] Stephen W. Lagakos. 2003. Clinical Trials and Rare Diseases. *New England Journal of Medicine* 348, 24 (2003), 2455–2456. https://doi.org/10.1056/NEJMe030024 PMID: 12802033.

[48] Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A. Terry, and Krzysztof Z. Gajos. 2016. Curiosity Killed the Cat, but Makes Crowdwork Better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (Santa Clara, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 4098–4110. https://doi.org/10.1145/2858036.2858144

[49] Qisheng Li, Krzysztof Z. Gajos, and Katharina Reinecke. 2018. Volunteer-Based Online Studies With Older Adults and People with Disabilities. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. ACM, New York, NY, USA, 229–241. https://doi.org/10.1145/3234695.3236360

[50] Qisheng Li, Sung Jun Joo, Jason D Yeatman, and Katharina Reinecke. 2020. Controlling for participants' Viewing Distance in Large-Scale, psychophysical online experiments Using a Virtual chinrest. *Scientific Reports* 10, 1 (2020), 1–11.

[51] I. Scott MacKenzie, Tatu Kauppinen, and Miika Silfverberg. 2001. Accuracy measures for evaluating computer pointing devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (Seattle, Washington, United States) *(CHI '01)*. ACM, New York, NY, USA, 9–16. https://doi.org/10.1145/365024.365028

[52] Michele Matarazzo, Teresa Arroyo-Gallego, Paloma Montero, Verónica Puertas-Martín, Ian Butterworth, Carlos S Mendoza, María J Ledesma-Carbayo, María José Catalán, José Antonio Molina, Félix Bermejo-Pareja, et al. 2019. Remote monitoring of treatment response in Parkinson's disease: the habit of typing on a computer. *Movement Disorders* 34, 10 (2019), 1488–1495.

[53] David E Meyer, Richard A Abrams, Sylvan Kornblum, and Charles E Wright. 1988. Optimality in human motor performance: Ideal control of rapid aimed movements. *Psychological Review* 95, 3 (1988), 340–370.

[54] Hugo Nicolau, Kyle Montague, Tiago Guerreiro, André Rodrigues, and Vicki L Hanson. 2017. Investigating laboratory and everyday typing performance of blind users. *ACM Transactions on Accessible Computing (TACCESS)* 10, 1 (2017), 1–26.

[55] Jukka-Pekka Onnela and Scott L Rauch. 2016. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* 41, 7 (2016), 1691–1696.

[56] Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. ACM, New York, NY, USA, 1364–1378. https://doi.org/10.1145/2675133.2675246

[57] Elena Rocco. 1998. Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 496–502.

[58] Cynthia Rothblum-Oviatt, Jennifer Wright, Maureen A Lefton-Greif, Sharon A McGrath-Morrow, Thomas O Crawford, and Howard M Lederman. 2016. Ataxia telangiectasia: a review. *Orphanet journal of rare diseases* 11, 1 (2016), 1–21.

[59] Remi M Sakia. 1992. The Box-Cox transformation technique: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)* 41, 2 (1992), 169–178.

[60] Jeremy D Schmahmann, Raquel Gardner, Jason MacMore, and Mark G Vangel. 2009. Development of a brief ataxia rating scale (BARS) based on a modified form of the ICARS. *Movement disorders* 24, 12 (2009), 1820–1828.

[61] Aasef G Shaikh, David S Zee, Allen S Mandir, Howard M Lederman, and Thomas O Crawford. 2013. Disorders of upper limb movements in ataxia-telangiectasia. *PLoS One* 8, 6 (2013), e67042.

[62] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[63] Neff Walker, David E Meyer, and John B Smelcer. 1993. Spatial and temporal characteristics of rapid cursor-positioning movements with electromechanical mice in human-computer interaction. *Human Factors* 35, 3 (1993), 431–458.

[64] Neff Walker, David A Philbin, and Arthur D Fisk. 1997. Age-related differences in movement control: adjusting submovement structure to optimize performance. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 52, 1 (January 1997), 40–53.

[65] Elke Warmerdam, Jeffrey M Hausdorff, Arash Atrsaei, Yuhan Zhou, Anat Mirelman, Kamiar Aminian, Alberto J Espay, Clint Hansen, Luc JW Evers, Andreas Keller, et al. 2020. Long-term unsupervised mobility assessment in movement disorders. *The Lancet Neurology* 19, 5 (2020), 462–470.

[66] Jerker Westin, Samira Ghiamati, Mevludin Memedi, Dag Nyholm, Anders Johansson, Mark Dougherty, and Torgny Groth. 2010. A new computer method for assessing drawing impairment in Parkinson's disease. *Journal of Neuroscience Methods* 190, 1 (2010), 143–148. https://doi.org/10.1016/j.jneumeth.2010.04.027

[67] Jacob O. Wobbrock and Krzysztof Z. Gajos. 2008. Goal Crossing with Mice and Trackballs for People with Motor Impairments: Performance, Submovements, and Design Directions. *ACM Trans. Access. Comput.* 1, 1 (May 2008), 1–37. https://doi.org/10.1145/1361203.1361207

[68] Jacob O. Wobbrock, Krzysztof Z. Gajos, Shaun K. Kane, and Gregg C. Vanderheiden. 2018. Ability-Based Design. *Commun. ACM* 61, 6 (may 2018), 62–71. https://doi.org/10.1145/3148051

## A MEASURES COMPUTED BY HEVELIUS

(1) **Movement time.** Complete movement time from target onset to the end of the successful click on the target.
(2) **Movement time variability.** Coefficient of variation of movement times in a block of trials.
(3) **Execution time.** Time from the first to the last mouse movement (excluding any movement that occurred while the mouse button was pressed – see Click slip).
(4) **Execution time without pauses.** Like execution time, but excludes pauses of 100ms or longer.
(5) **Execution time variability.** Coefficient of variation of execution times in a block of trials.
(6) **Execution time variability (without pauses).** Coefficient of variation of execution times (without pauses) in a block of trials.
(7) **Peak speed.** The maximum (smoothed) speed recorded during a movement.
(8) **Peak speed variability.** Coefficient of variation of peak speeds in a block of trials.
(9) **Peak acceleration.** The maximum (smoothed) acceleration recorded during a movement.
(10) **Peak acceleration variability.** Coefficient of variation of peak accelerations in a block of trials.
(11) **Distance from target center at end of main submovement.** The 2D distance from the mouse pointer location at the end of the main submovement to the target center.
(12) **Fraction of remaining distance to the target center covered in main submovement.** The fraction of the remaining distance along the task axis covered during the main submovement. The value of this measure can be greater than 1 if the person overshoots the target.
(13) **Maximum deviation from task axis.** The maximum distance of the mouse pointer from the task axis during a movement.

(14) **Movement variability.** The standard deviation of the distance of the actual path from the task axis [51].

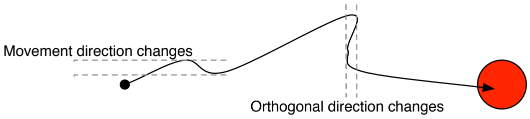
(15) **Movement error.** The average absolute distance of the mouse pointer from the task axis. In other words, this measure captures, at the gross level, how far the pointer trajectory was from a straight line [51].

(16) **Movement offset.** The average (non-absolute) distance of the mouse pointer from the task axis. A large magnitude of movement offset indicates that the movement trajectory falls mostly to one side of the task axis or the other. A movement with a large movement error may still have a small movement offset if the path of the movement deviates first to one side of the movement axis and then to the other [51].

(17) **Task axis crossings.** The number of times the mouse pointer crossed the task axis during the movement [51].

(18) **Target re-entries.** The number of times the mouse pointer leaves the target and then re-enters it before the start of the click [51].

(19) **Movement direction changes.** The number of times the movement component orthogonal to the task axis changes sign (illustrated below) [51].



(20) **Orthogonal direction changes.** The number of times the movement component parallel to the task axis changes sign (illustrated above) [51].

(21) **Main submovement.** The submovement with the highest peak speed.

(22) **Verification time.** The time interval between the end of a movement inside a target and the beginning of the click (i.e., the time when the mouse button was pressed).

(23) **Verification time variability.** Standard deviation of verification times in a block of trials.

(24) **Click duration.** The time between mouse button press and release during the correct click on the target.

(25) **Click duration variability.** Standard deviation of click durations in a block of trials.

(26) **Click slip.** Distance between the point where the mouse button was pressed down and where it was released during click on the target.

(27) **Noise-to-force ratio.** The standard deviation (computed over all trials in a block) of the distance from the target center at the end of the first submovement divided by mean of peak accelerations [64].

(28) **Normalized jerk.** A dimensionless measure computed as

$$\text{normalized jerk} = \frac{(ET)^3}{v_{max}^2} \int_t \left(\frac{da}{dt}\right)^2 \mathrm{d}t$$

where $\frac{da}{dt}$ is the jerk, $ET$ is the execution time (excluding pauses) and $v_{max}$ is the peak speed during the movement. While some researchers use mean speed rather than peak speed [27], others (e.g., [3]) used peak speed instead. We found that normalized jerk computed with peak speed correlated less with the index of difficulty of a movement than normalized jerk computed with mean speed.

While numerous jerk-based measures have been used in prior research (e.g., square integrated jerk [67]), normalized jerk was designed to be minimally correlated with task properties (target size, distance to the target) [27].

(29) **Normalized jerk without pauses.** Like normalized jerk, but excludes parts of the movement when the mouse pointer was paused for 100ms or longer.
(30) **Fraction of the main submovement spent accelerating.** The fraction of the time from the start of the submovement to the time when acceleration reached its peak value divided by the overall duration of the submovement.
(31) **Number of pauses.** Number of pauses of 100ms or longer.
(32) **Duration of the longest pause.** Duration of the longest pause of 100ms or longer. If no such pause occurred, 0ms is recorded for this measure.

## B PARAMETERS OF THE UPDATED MODEL FOR ESTIMATING BARS DOMINANT ARM SEVERITY SCORES

Table 5 shows the Hevelius measures and the corresponding weights used in the regression model for estimating the dominant arm component of the BARS score. The dominant arm component of the BARS score ranges from 0 (no impairment) to 4. The output of the model was restricted to produce values within that range.

Table 5. Hevelius measures and the corresponding weights used in the linear regression model for estimating the dominant arm component of the BARS score.

| Measure name | Weight |
|---|---|
| Intercept | 0.0679 |
| Movement time | 0.1151 |
| Click duration | 0.1042 |
| Main submovement | 0.0820 |
| Movement direction changes | 0.0632 |
| Fraction of the distance to the target center covered during main submovement | 0.0427 |
| Number of pauses | 0.0382 |
| Execution time (w/o pauses) variability | 0.0268 |
| Verification time | -0.0265 |
| Execution time | 0.0155 |
| Target reentries | 0.0059 |
| Click slip | 0.0010 |