Validation of an Automated Scoring Program for a Digital Complex Figure Copy Task within Healthy Ageing and Stroke

Sam S. Webb*[1], Margaret Jane Moore*[1], Anna Yamshchikova[1], Valeska Kozik[1], Mihaela D. Duta[1], Irina Voiculescu[1] & Nele Demeyere[1]

(*joint first authors)

[1]Department of Experimental Psychology, University of Oxford

Accepted for publication in Neuropsychology, DOI: http://dx.doi.org/10.1037/neu0000748

Author Note

Sam S. Webb https://orcid.org/0000-0002-0029-4665

Margaret Jane Moore https://orcid.org/0000-0001-6914-5978

Anna Yamshchikova https://orcid.org/0000-0002-0920-3301

Valeska Kozik https://orcid.org/0000-0002-7895-6788

Mihaela D. Duta https://orcid.org/0000-0002-0435-571X

Irina Voiculescu https://orcid.org/ 0000-0002-9104-8012

Nele Demeyere https://orcid.org/0000-0003-0416-5147

We have no conflict of interest to disclose.

Correspondence concerning this article should be addressed to Nele Demeyere Department of Experimental Psychology, New Radcliffe House, Radcliffe Observatory Quarter, Oxford, OX2 6AE. Tel: 01865 271340. Email: nele.demeyere@psy.ox.ac.uk

This project was presented at the UK Stroke Forum, December 2020.

Abstract

Objective

Complex Figure Copy Tasks are one of the most commonly employed neuropsychological tests. However, manual scoring of this test is time-consuming, requires training, and can then still be inconsistent between different examiners. We aimed to develop and evaluate a novel, automated method for scoring a tablet-based Figure Copy Task.

Method

A cohort of 261 healthy adults and 203 stroke survivors completed the digital Oxford Cognitive Screen – Plus Figure Copy Task. Responses were independently scored by two trained human raters and by a novel automated scoring program.

Results

Overall, the Automated Scoring Program was able to reliably extract and identify the separate figure elements (average sensitivity and specificity of 92.10% and 90.20% respectively) and assigned total scores which agreed well with manual scores (ICC = .83). Receiver Operating Curve analysis demonstrated that, compared to overall impairment categorisations based on manual scores, the Automated Scoring Program had an overall sensitivity and specificity of 80% and 93.40% respectively (AUC = 86.70%). Automated total scores also reliably distinguished between different clinical impairment groups with sub-acute stroke survivors scoring significantly worse than longer term survivors, which in turn scored worse than neurologically healthy adults.

Conclusions

These results demonstrate that the novel automated scoring algorithm was able to reliably extract and accurately score Figure Copy Task data, even in cases where drawings were highly distorted due to comorbid fine-motor deficits. This represents a significant advancement as this novel technology can be employed to produce immediate, unbiased, and reproducible scores for Figure Copy Task responses in clinical and research environments.

Keywords: neuropsychology, stroke, validation, automated scoring, complex figure copy

Key points

Question

We aimed to develop and evaluate a novel, automated method for scoring a tablet-based Figure Copy Task

Findings

The novel Automated Scoring Program was able to reliably extract and accurately score Figure Copy Task data, even in cases where drawings were highly distorted due to comorbid fine-motor deficits.

Importance

This represents a significant advancement as this novel technology can be employed to produce immediate, unbiased, and reproducible scores for Figure Copy Task responses in clinical and research environments

Next steps

Trialing the Automated Scoring Program in clinical environments

Validation of an Automated Scoring Program for a Digital Complex Figure Copy Task within Healthy Ageing and Stroke

The administration of neuropsychological tests is a key component of establishing brain-behavior relationships (Crawford et al., 1992; Ellis & Young, 2013). However, comparisons which employ these metrics can be limited by the quality of scoring of these neuropsychological tests. For example, tests which require subjective examiner judgements may introduce potentially confounding noise into neuropsychological analyses (Barker et al., 2011; Franzen, 2000; Moore et al., 2019; Watkins, 2017). Inter-rater reliability traditionally is improved by implementing extensive training courses, employing exhaustive or requiring agreement across multiple independent raters or tests (Franzen, 2000; Huygelier et al., 2020). However, more demanding scoring procedures often are prohibitively time-consuming and can lead to studies opting to rely on small, selected samples rather than larger, generalizable patient cohorts, or similarly to only complete limited cognitive measures (e.g., MMSE Folstein et al., 1983) which reduce the informational richness. For these reasons, identifying new methods for efficiently improving scoring consistency on clinically feasible measures is critically important for improving both the scope and reliability of neuropsychological investigations. Here, we focus on validating this approach in a specific, prominently studied, clinical cohort of stroke survivors, as an example group where these automated scoring measures may improve methods to further elucidate specific aspects of domain-specific cognitive impairments in Complex figure copy and recall.

The Figure Copy test is one of the most commonly employed neuropsychological assessment methods used to evaluate visuospatial constructional ability and nonverbal memory in clinical environments (Shin et al., 2006). In traditional versions of this test, participants complete two drawings of a composite geometric shape. First, participants are presented with a target image and are asked to copy it from sight. Next, the target figure is removed and participants are asked to reproduce it from memory (Demeyere et al., in press; Schreiber et al., 1999). The Rey-Osterrieth Complex Figure Test (ROCFT; Somerville et al., 2000) is the most well-known figure copy test, though many variations, including computerised versions (e.g., Demeyere et al., in press; Humphreys et al., 2017; Schreiber et al., 1999; Taylor, 1969), are in use.

Successful completion of any figure copy task requires participants to coordinate fine-motor movements, employ visuospatial perception, maintain visual images in working memory, and effectively plan and organise their responses (Shin et al., 2006). The Figure Copy Task has been found to act as a reliably metric of a wide range of cognitive functions, and is therefore useful for establishing a diverse range of brain-behavior relationships. Chechlacz et al. (2014) conducted a voxel-lesion symptom mapping study aiming to identify the neural correlates of a range of deficits captured by performance in a figure copy task. Analysis of this single behavioral assessment yielded significant and distinct neural correlates associated with general poor performance, lateralized omissions, spatial positioning errors, global feature impairment, and local feature impairment (Chechlacz et al., 2014). Similarly, (Chen et al., 2016) conducted a lesion mapping study investigating the correlates of principal component analysis-derived factors underlying figure copy performance. This investigation identified brain regions associated with high-level motor control, visuo-motor transformation, and multistep object use using only behavioral data from a figure copy task. This wide range of assessed cognitive functions makes the figure copy task an extremely valuable tool both for clinical diagnostic purposes and for research aiming to establish brain-behavior relationships.

The Figure Copy Task is comparatively simple to complete and while assessing a diverse range of functions. These advantages mean that this task is frequently employed within clinical neuropsychological evaluations. A survey conducted by Rabin et al. (2016) found that the ROCFT was the eighth most popular single neuropsychological assessment employed by a sample of 512 North American neuropsychologists, with 7.6% reporting using this test (Rabin et al., 2016). Previous research has suggested that Figure Copy Task performance can effectively distinguish between various clinical populations (Alladi et al., 2006; Demeyere et al., in press; Freeman et al., 2000). For example, Freeman et al. (2000) administered the Rey-Osterrieth Complex Figure test to a cohort of Alzheimer's disease, ischemic vascular dementia, and Parkinson's disease patients. This investigation identified significant differences in performance with patients with Alzheimer's disease performing significantly worse than patients diagnosed with vascular dementia or Parkinson's Disease (Freeman et al., 2000). These findings suggest that patients' Figure Copy Task scores may provide clinically relevant information which can be employed to inform diagnoses.

Patient performance on Figure Copy task is generally scored manually. For example, examiners score performance on the Oxford Cognitive Screen – Plus (OCS-Plus) figure copy task by reporting the presence, accuracy, and position of each individual figure element independently (Demeyere et al., in press). However, this scoring method is time-consuming, requires training, and is ultimately reliant on subjective examiner impressions. Individual examiners may disagree on which drawn line represents which element, especially in cases where a patient has committed many errors. A significant amount of training is required to ensure high agreement. This reliance on subjective examiner judgements inevitably introduces human biases into Figure Copy scores. Relying on subjective interpretations of objective criteria can result in systematic scoring biases, potentially precluding the validity of large-scale comparisons involving Figure Copy Test data, especially in cases where multiple independent examiners are involved. Automated algorithms have been repeatedly demonstrated to be able to perform many diagnostic and classification tasks with greater sensitivity and specificity than human experts (Dawes et al., 1989; Meehl, 1954).

For this reason, several automated tools have been developed to quantify performance on neuropsychological tests. Chen et al. (2020) developed a deep-learning based automated scoring tool for the Clock Drawing Task, a common component of dementia screening batteries (Agrell & Dehlin, 1998; Pinto & Peters, 2009). This investigation compared algorithmic and expert assigned scores in a cohort of 1315 outpatients and concluded that the algorithm exhibited a comparative scoring accuracy of 98.54% (Chen et al., 2020). Similarly, Moetesum et al. (2015) applied an automated approach to assessing performance on the Bender Gestalt Test (Koppitz, 1964) within a sample of 18 healthy adults. The performance of this algorithm varied dramatically depending on the specific gestalt component being assessed (range = 6/18 (overlap) and 18/18 (rotation)) (Moetesum et al., 2015).

Two figure-copy specific automated scoring algorithms have been developed. First, Canham et al. (2000) developed an automated scoring software for the commonly used Rey-Osterrieth Complex Figure test. In this task, responses are generally manually scored by categorising each of the target figure’s 18 elements according to whether or not they are present, accurately drawn, and correctly placed within the response figure. Canham et al.'s (2000) automated software matched these scoring criteria by first identifying distorted areas of patient drawings, then locating and grading basic geometric shapes while employing unary metrics to remove unsuitable features from patient drawings. This method was found to perform well on real patient data with 75% of features being within 5% of the manually assigned scores and 98.6% within 10% (Canham et al., 2000). Second, the most recent, “state-of-the-art” figure copy scoring tool was designed by Vogt et al. (2019), which demonstrated a .88 Pearson correlation with human ratings of Rey-Osterrieth Complex Figure performance. While this performance in near the documented human inter-rater agreement (.94), equivalence testing revealed that these scoring methods did not produce strictly equivalent total scores. However, these algorithms were designed specifically to score data from the Rey-Osterrieth Complex Figure test and do not generalise to other commonly used Figure Copy Tests.

The purpose of the present investigation is to develop an automated scoring tool to score the OCS-Plus (Demeyere et al., in press) Figure Copy Task. This project aims to evaluate the efficacy of this automated scoring tool by comparing automated versus manually assigned scores and identifying potential sources of systematic disagreement. The utility of this automated software for distinguishing between different clinical populations is also explored. Ultimately, this project aims to deliver a robust automated clinical scoring tool to deliver immediate scoring and evaluation of individual performance on the OCS-Plus Figure Copy Task.

Methods

Participants

A cohort of 261 neurologically healthy adults were recruited as well as 203 stroke survivors who completed the Figure Copy task within the OCS-Plus Tablet Screening Project (REC reference: 18/SC/0044, IRAS project ID: 241571). Of the stroke survivors 49 were tested on the Figure Copy test within 6 months of their stroke (termed sub-acute stroke participants) and 154 stroke survivors were tested on the Figure Copy test on or after 6 months post-stroke (termed chronic stroke participants).

All healthy adult participants were recruited through convenience sampling as part of the OCS-Plus validation project (Demeyere et al., in press) from an existing pool of older healthy ageing research volunteers (*Anonymous* University, MSD-IDREC-C1-2013-209). Healthy adult participants were included in the OCS-Plus project if they were able to provide informed consent, had sufficient English language proficiency to comprehend instructions, were at least 18 years old, and were able to remain alert for at least 20 minutes. The exclusion criteria included inability for the participant to consent to take part, insufficient English language proficiency, and inability to stay alert for 20 minutes to do the task.

Table 1

Summary demographics of the samples (49 = sub-acute stroke, 154 chronic stroke, 261 healthy adults)

	All healthy adults (*n*= 261)	Sub-Acute stroke survivors (*n*=49)	Chronic stroke survivors (*n*=154)	HA vs. AS	HA vs. CS	AS vs. CS
Average Age (SD)	60.39 (15.76)	70.75 (15.45)	72.17 (13.67)	0.66*	0.78*	-0.10
Education: Average yrs (SD)	15.69 (4.09)	13.23 (4.38)	12.5 (3.46)	-0.60*	-0.81*	0.20
Handedness % right	89.66%	94.7%	68.18%	.05	.11	.08
Sex % female	53.64%	57%	44.81%	.04	0	.07
Stroke side L=left R= Right B= bilateral		L= 36.73% R = 46.94% B = 6.12%	L= 33.77% R = 38.96% B = 14.29%			.12
Stroke Type Ischemic (IS) Haemorrhagic (ICH) TIA		IS= 77.55%, ICH= 18.37% TIA = 2.04%	IS = 74.03% ICH = 14.94% TIA = 2.6%			.12
Average Days Since Stroke (SD, Range)		72.53 (82.37, 1- 181)	451.63 (446.07, 182 - 2658)			-0.99*

Average Lesion Volume (SD) in cm^3	21.03 (36.07)	39.30 (51.04)	-0.38*

Note. HA refers to healthy adults, AS to sub-acute stroke, and CS to chronic stroke groups. For age, education, days since stroke, and lesion volume, groups were compared using independent t-tests and we report the Cohen's *d* effect size. For handedness, sex, stroke side, stroke type, we used chi squared analysis and report Cramer's V effect size. * refers to significance below .05.

We collected additional measures from clinical notes including the Barthel Index (Mahoney & Barthel, 1965) and the Oxford Cognitive Screen (Demeyere et al., 2015) to measure functional ability and domain-specific cognitive impairment. As part of the 6-month follow up protocol for the overarching study, we collected data on the Hospital Anxiety and Depression Scale (Zigmon & Snaith, 1983), to measure anxiety and depression, the Stroke Impact Scale (Duncan et al., 2002) to measure the domain-specific impact of stroke, and the Quality of Life Scale (Al-Janabi, Flynn, & Coast, 2012) to assess the quality of life of the participants post-stroke.

Table 2

Additional tests of mood, cognitive impairment, stroke severity, and the impact of stroke, for most of the stroke survivor sample (Sub-Acute stroke = 49, chronic stroke = 154)

		Sub-Acute				Chronic			
Test battery	Measure	n	missing (%)	*M*	*SD*	n	missing (%)	*M*	*SD*
	Barthel index					124	19	16	5
Oxford Cognitive Screen	Number of domains impaired	49	0	3	2	148	4	3	2
Hospital Anxiety Scale	Anxiety					135	12	6	4
	Depression					134	13	5	4
Stroke Impact Scale	Total	49	0	115	142	154	0	242	111
	Strength	46	6	6	8	139	10	12	6
	Hand	46	6	6	9	141	8	15	9
	ADL	46	6	15	19	143	7	35	15
	IADL	47	4	12	15	141	8	25	12
	Mobility	47	4	14	17	143	7	31	13
	Communication	46	6	13	16	142	8	28	9
	Emotion	47	4	14	17	141	8	29	10
	Memory	44	10	12	15	142	8	26	9
	Quality of Life scale					128	17	15	3

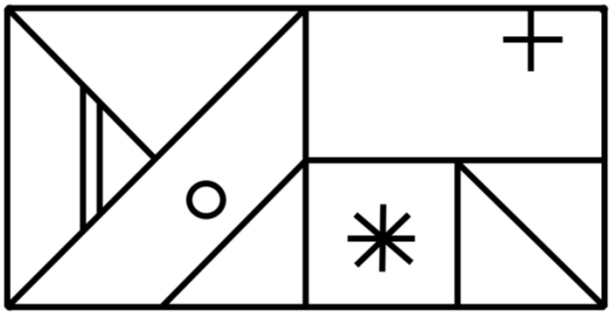
Note. Sub-Acute and Chronic refer to when the Figure Copy test was administered, so either before 6 months post-stroke (termed sub-acute) or greater or equal to 6 months post-stroke. ADL refers to activities of daily living, IADL refers to instrumental activities of daily living.

The OCS-Plus Figure Copy Task and Manual Scoring Criteria

The OCS-Plus is a tablet-based cognitive screening tool designed to briefly assess cognitive impairments within clinical and sub-clinical populations using fine-grained measures (Demeyere et al., in press). The OCS-Plus version used in this investigation was created in MATLAB 2014b and was run on a Microsoft Surface Pro computer tablet (Windows 10 Pro, version 1511). The OCS-Plus begins with a small practice to ensure even those with limited experience with computer-tablet technology can complete tasks accurately, this practice involves tapping a shape in the centre of the screen, and drawing a line between two small dots. The OCS-Plus includes a computerized Figure Copy Task which is designed to be inclusive for severely impaired patients, including a simple, multi-element target figure. In this task, participants are asked to copy a composite geometric shape (Figure 1) once from sight and again from memory, immediately following completion of the copy condition. Participants are not informed that they will be asked to remember the figure until the beginning of the memory condition. Participants are instructed to complete their drawing using a tablet stylus within a marked area underneath the target figure. Participants are allowed unlimited time to complete each of these drawing tasks.

Figure 1

The target figure in the OCS-Plus Figure Copy Task. See Figure 2 for individual figure element definitions.



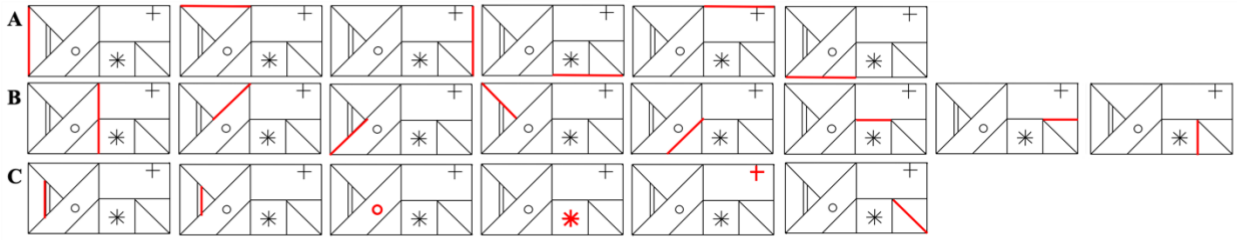
Note. Figure available at https://osf.io/gkbvd/ under a CC-BY4.0 license.

The Figure Copy Task records performance in terms of coordinates and timeline, allowing full, detailed reconstruction of the drawing process. Each completed drawing is assigned a total score out of 60 with each of the individual 20 figure elements being scored independently according to three independent criteria: *presence*, *accuracy*, and *position* (Figure 2). An element is scored as *present* if it has been drawn anywhere in the response figure. Perseverative responses are not quantitatively penalized but are noted by the examiner. Elements are marked as *accurate* if they are drawn with reasonable accuracy as could be expected from a person with typical drawing ability. Reasonable allowances are made to account for the use of a tablet computer stylus on the relatively slippery screen surface and comorbid age-related fine motor impairments (e.g., arthritis). For example, slight inaccuracies in line joining as well as obvious attempts to correct such errors (e.g., doubling up a line to ensure that it is straight) are not penalised. Finally, element *position* is marked as correct if an element's location is reasonably accurate. As in accuracy scores, allowances are made to account for tablet usage and age-related fine motor impairments. Scorers are instructed to only penalise each drawing position error once and to disregard cases in which position errors within one element have led to placement errors within neighbouring elements. These criteria are used to assign a score out of three for each individual element shown in Figure 2, and these element scores are summed to produce a total score. This scoring procedure is repeated for the copy and recall drawing condition.

A full scoring manual detailing the exact instructions given to scorers is openly available on the Open Science Framework (Foster & Deardorff, 2017; https://osf.io/9dwpv/). Human raters completed approximately two hours of training with the manual to complete the manual ratings. The average time required to manually score a figure copy response varied between 1 and 5 minutes, depending on the degree of distortion and error present within the response drawing. The automated scoring programme requires less than 5 seconds to score a drawing, implementation of automated scoring can be expected to save between 2-10 minutes per participant (2 drawings each). Note, we automatically scored full points (18 points) for border elements if the participant had used the drawing area border as the figure border (i.e., had drawn no border elements). This approach was adopted in order to avoid penalising participants who used the outer border of the rectangular drawing area as the figure border. Given that this error pattern occurred in both healthy adult participants (*n*=23 in the copy condition, *n*=27 in the recall condition) and patients (*n*=14 in the copy condition and *n*= 11 in the recall condition), we judged this to not have represented a clinical deficit, such as closing-in behaviour, and instead attributed these errors to the presentation of the space indicating where to draw on the tablet being too similar in size and shape to the drawing, along with potential misunderstanding of instructions. Given the small number of patient responses (14 at maximum, and more for the healthy adults) this scoring rationale did not significantly impact the results of the conducted analyses.

Figure 2

Each of the stimulus elements which are independently scored.



Note. Elements are divided into three sections: A) border lines, B) internal dividers, and C) the detail components. Each element is given a score out of three with one point being awarded based on presence, accuracy, and position independently. If viewed in black and white, each element on the figure is highlighted individually in red in each mini figure. Figure available at https://osf.io/x2ks7/ under a CC-BY4.0 license.

Automated Scoring Program

The Automated Scoring Program created in this project was developed in Python 3.7 and employs functions from the packages SciPy (Jones et al., 2001), Shapely (Gillies, 2015), Kivy (Virbel et al., 2011), and PyLaTeX (Fennema, 2014). This program employs output variables created by the OCS-Plus software including (x,y) coordinates of patient responses, time stamps, and final drawing images. Before scoring each element, this software first pre-processes this data in six sequential steps: noise removal, normalisation, circle identification, line segmentation/extraction, star and cross identification, and line element identification.

First, in noise removal, all pen strokes totalling fewer than five pixels are removed, as these responses represent very small marks which were most likely created by accidentally touching the pen to the tablet. Similarly, all elements which are abnormally distant from other elements are removed, as these marks are unlikely to be a part of a participant's intended response. Abnormal distance is determined via calculating the centroids for each element, and then use k-dimensional trees (Maneewongvatana & Mount, 2002) to find nearest neighbours for each of the centres within the distance r, such that $r = \frac{1}{2}min(fig_h, fig_w)$, where h is height of figure and w is width. Second, participant drawings are normalised. This step is essential due to the large variance in participant response sizes, orientation angles, and positions within the allocated drawing response area. Normalisation is conducted by translating each drawing to be positioned with the bottom left-most point at coordinate (0,0) then scaling the x and y axis to match the dimensions of the target figure.

In the third step, circular elements are identified within the normalised response drawing. Circles are defined as a continuous path which meets the criteria detailed in Figure 2. The values of these parameter cut-offs were adjusted to the values which optimise overall performance. Next, line segmentation is performed using the Ramer-Douglas-Peucker algorithm which processes a series of points on a single curve and outputs a simplified element path composed of straight lines (Douglas & Peucker, 1973). Vector calculations are then used to determine the angle between multiple lines on each simplified curve, to identify turning points, and to subsequently split simplified lines into individual figure elements.

In step five, star and cross figure elements are identified by finding all sets of lines composed of intersecting paths where the length of each line is less than half of the drawing's total height and individual line lengths are within the third quartile plus 1.5 of the interquartile range of each of the intersecting lines. Line sets of three or more lines where the smallest angle between lines is greater than or equal to 30 degrees are defined as stars and sets of 2 or more lines of which the smallest angle between lines is greater than or equal to a threshold, empirically determined at 36 degrees are defined as crosses. Finally, in the last step, line elements of the response figure are identified. The orientation of each remaining unclassified figure element is determined as either vertical, horizontal, right, or left slanted by calculating the angles between simplified lines and the normalised x-axis. Euclidian distance calculations (Deza & Deza, 2009) are then used to match each drawn line to its corresponding element in the target figure.

Once this six-step pre-processing is completed, response drawing total scores are assigned automatically. As in manual scoring, each element is assessed based on presence, accuracy, and position. The Automated Scoring Program marks an element as present if it has been identified in the pre-processing steps described above. Accuracy scoring criteria differ based on the element being assessed. For components such as circles, stars, and crosses to be successfully identified by the pre-processing, they must be drawn with a reasonable degree of accuracy. For this reason, if a circle, star, or cross is marked as being present, it is also scored as being accurate. The accuracy of linear elements is scored by calculating the best fit line of the element via linear regression.

The distance between a drawn point and a target point in 2D space is calculated as the absolute difference between their respective x and y-coordinates. Linear elements are scored as accurate if the maximum distance from any point of the target element to the best fit regression line is less than 10, the length of the best fit line is greater than or equal to 70% of the target path length, and the angle between the best fit line and target line segment is less than 10 degrees. If two line segments, which are defined as separate in the figure template, are drawn as a continuous line in the participant’s drawing, the algorithm is able to split the drawn line segment in order to assess fit of the separated line segments to the original template as to avoid underscoring presence.

Finally, element position is scored by comparing the location of drawn paths to the location of the corresponding element within the target figure. The algorithm assigns each drawn linear element to its corresponding target element, if it has the same orientation and the distance between the elements is less than 20% of the total drawn figure height. As these position criteria have to be met in order to identify a line, such a line is automatically scored as being in the correct position. The detail elements star, cross, and circle are scored as positioned correctly if their distance from the target location is less than 50% of the drawn figure height. Similarly, to manual scoring, the automatic scoring program scores full border elements points if the all border elements are not present. This scoring process results in a total score out of 60 points for each response drawing. Full details on the design and implementation of this Automated Scoring Program can be found in the original masters dissertation which details the Program (Yamshchikova, 2019). The Figure Copy software can be downloaded for Academic Use from Oxford University Innovation Software Store (https://process.innovation.ox.ac.uk/software/).

Data analysis

The manual scoring data included in this investigation was completed independently by *Anon author 1* (rater 1) and *Anon author 2* (rater 2). Both raters were trained to score drawings and both scored all 928 responses included in this investigation. During scoring, all figures were randomized and anonymized so that raters were blind to drawing condition, participant group, and identity. First, the degree of agreement between human rater scores was assessed. Given that figure copy total scores represent an aggregate measure which may not accurately capture inter-element variation, these analyses were conducted on total scores and on an element-wise basis. Agreement was measured in two ways. First, summed scores were compared using an intraclass correlation coefficient (ICC; model, ICC1: i.e., single scores, random raters), which measures the ratio of true variance divided by true variance plus error variance (Koo & Mae, 2016) and ranges from 0 to 1. Cohen's kappa reliability statistic was used for binary data such awarding a presence, accuracy, or position score or not and is scaled as a standardised correlation coefficient to enable cross-study interpretation (McHugh, 2012). This investigation employs the ICC reliability benchmarks proposed by Koo and Mae (2016): <=.50 = poor reliability; >.50 - <=.75 = moderate reliability; >.75 - <=.90 = good reliability; >.90 = excellent reliability. All Cohen's kappa calculations employ the agreement benchmarks defined by McHugh (2012): 0-.20 = no agreement; .21-.39 = minimal agreement; .40-.59 = weak agreement; .60-.79 = moderate; .80-.90 = strong; >.90 = almost perfect.

Next, the agreement between the Automated Scoring Program and aggregate human scores was determined. Element-wise sensitivity (True Positives / True Positives + False Negatives) and specificity (True negatives / True Negatives + False Positives) was calculated. In these calculations, False Negatives represented cases in which an element was identified by manual scoring, but not by the automated program. Conversely, False Positives represented cases where an element was identified by the automated program, but not by human raters. Sensitivity analysis is usually used in the case of determining whether a test correctly identifies a specific group of cases from another, in our case presence of an element or not. The benchmark for interpretation is that sensitivity + specificity should be close to or above 1.50 (or 150-200 when as a percentage as reported), where a value of 1 reflects an uninformative test and a value of 2 represents a perfect test (Power, Fell & Wright, 2013).

We also examined how the Automated Scoring Program resolved cases in which the raters did not assign identical scores. Next, a qualitative analysis of cases in which the automated program was and was not able to extract meaningful scores was conducted. Finally, the known-group discriminability of total scores assigned by the Automated Scoring Program was examined.

Statistical analyses were conducted in R (version 3.5.1, 2018-07-02), R Core Team, 2018), the data and analyses scripts used to generate this manuscript are openly available (https://osf.io/3k6gs/). We used the following packages for statistical analyses and visualisation *ggplot2* (version 3.3.2; Wickham, 2016), *cowplot* (version 1.1.0; Wilke, 2019), *psych* (version 1.8.12; Revelle, 2018), *irr* (version 0.84.1; Gamer et al., 2019), *pROC* (version 1.16.2; Robin et al., 2011), *rcompanion* (version 2.3.7; Mangiafico, 2019), *rstatix* (version 0.4.0; Kassambara, 2020), and *plyr* (version 1.8.5; Wickham, 2011).

Results

Human inter-rater reliability

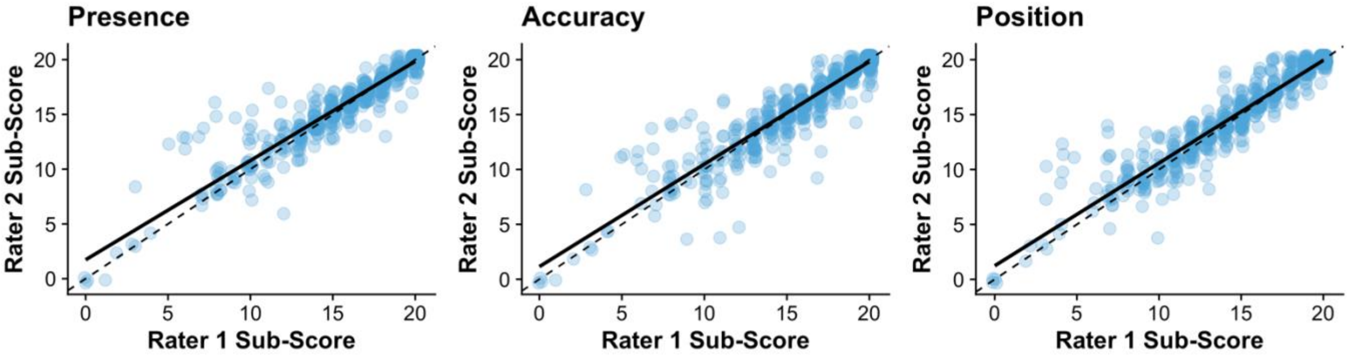
The average total score assigned by human raters was 57.72 (*SD*=6.09) for copy condition drawings and 44.4 (*SD*= 10.69) for recall condition responses. Raters exhibited a high degree of agreement between assigned total scores, with a cumulative interclass correlation (ICC) of .97, *F*(927,927)=58.40, *p* <.001, 95% CI [.96-.97]. This close agreement was present within both the copy (ICC =.97, *F*(463,463)=57.56, *p* <.001, 95% CI [96-.97]) and recall (ICC = .94, *F*(463,463)=33.04, *p* <.001, 95% CI [93-.95]) condition drawing scores.

Of the 55680 elements scored, only 3.64% were assigned conflicting element sub-scores by the assessors. Of all elements, raters disagreed on position scores most frequently (1.30%), followed by accuracy scores (1.26%), and then presence (1.08%). See Figure 3. Raters were found to disagree on more recall condition elements (5.51%) than copy condition elements

(1.78%). This difference is likely due to the comparatively greater quality variation present within delayed recall drawing responses (recall variance = 126.63, copy variance = 17.41).

Figure 3

Illustrates the between-rater associations for element-wise sub-scores



Note. Between-rater total score comparisons across both copy and recall condition figure copy drawings (*N*=928) demonstrated a high degree of agreement across total accuracy, position, and presence scores (ICCs = .95, .96, and .96 respectively). The dashed line represents perfect correlation (slope =1, intercept = 0) in order to demonstrate deviation of agreement, and solid line reflecting best fit line. Figure available at https://osf.io/2qwfn/ under a CC-BY4.0 license.

Next, elements which caused the highest degree of disagreement between the raters were identified. The most frequent element to be disagreed upon across all subs-scores was the middle bottom right interior divider slanted line (element 11, see Table 4) where the human raters disagreed on all three sub-scores a total of 37 / 928 times (3.99%). The small left vertical interior divider line (element 12) had the highest number of two sub-score disagreements (4.42%, *n* = 41), with position representing the most commonly disputed sub-score. Finally, the circle (element 14) had the highest number of cases in which human raters differed within a single sub-score (6.14%, *n* = 57). This disagreement primarily impacted position scores.

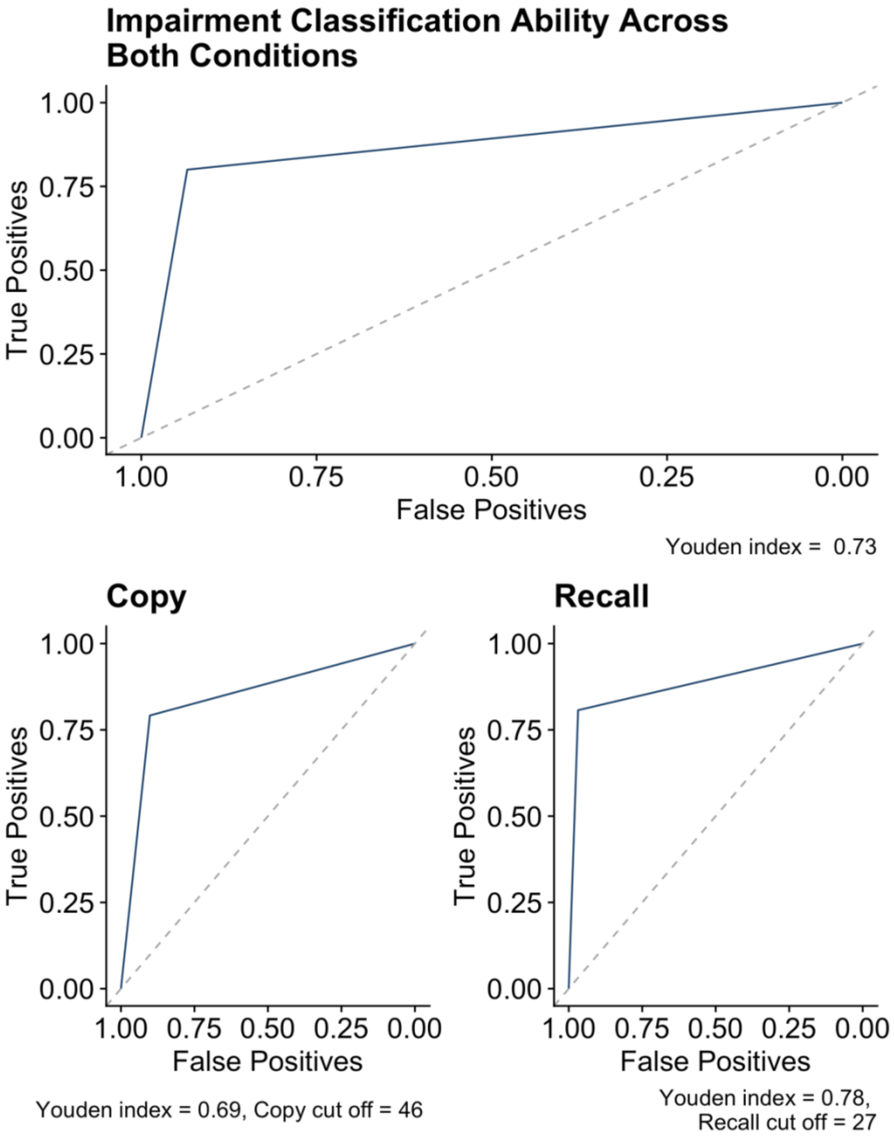
Automated Scoring Program versus Human Raters

The comparative accuracy of the automated figure copy scores was evaluated against the manually assigned scores. For the element-wise analyses, only element scores where both raters agreed (96.36% of all scores) were included in these analyses. For total score comparisons, we averaged the two raters total scores. This procedure was adopted to ensure the Automated Scoring Program was able to accurately score figures versus agreed-upon scores before moving on to more complex cases. Overall, the scores assigned by the automated program and raters exhibited a high degree of agreement both in terms of total score (ICC = .83 *F*(927,927)=21.90, *p* <.001, 95% CI [.23-.93]) and element scores (Cohen's *k* = .63, 95% CI [.63-.63], *p*<.001). The same was true for the recall condition (total score ICC = .83, *F*(463,463)=22.69, *p* <.001, 95% CI [.23-.94], element-wise agreement Cohen's *k* =.71, 95% CI [.71-.72], *p*<.001), and copy condition (total score ICC = .58, *F*(463,463)=7.20, *p* <.001, 95% CI [.01-.81], element-wise agreement Cohen's *k* = .28, 95% CI [.28-.30], *p*<.001).

A further way to compare the Automated Scoring Program to the human raters scoring, is to compare whether the same participants are identified as impaired on either scoring version. Receiver operating characteristic (ROC) analyses were conducted to compare total score binarized impairment categorizations (i.e., less than 2*SD*s below the mean) of the automated assigned scores to those based on the standard manual scoring and cut-offs in copy and recall conditions. In this way we directly compared the impairment classification between manually and automatically derived scores, rather than trying to determine presence of a stroke event. When compared to impairment categorisations made based on manual scores overall (i.e., across both copy and recall conditions), the Automated Scoring Program was found to have a total sensitivity of 80%, a specificity of 93.44%, and an AUC of 86.72%, 95 CIs [82.72-90.66%], Youden index = .73. The sensitivity and specificity were similarly high in the copy and recall condition, with a slightly lower Youden index in the copy condition (sensitivity = $79.13\%_{\text{copy}}$ & $80.70\%_{\text{recall}}$, specificity = $90.14\%_{\text{copy}}$ & $96.81\%_{\text{recall}}$, Youden index = $.69_{\text{copy}}$ & $.78_{\text{recall}}$). When overall sensitivity and specificity are summed, we get a value of 173.44% or 1.73 in raw units, meaning that our test had above excellent ability determine impairment classification compared to manual scores (Power et al., 2013). Table 3 summarises the average scores attained by each sample group per copy and recall condition, and presents group specific sensitivity and specificity statistics.

Figure 4

ROC curve illustrating sensitivity/specificity of the Automated Scoring Program binarised impairment in comparison to averaged rater scores binarised impairment.



Note. There were separate cut offs for recall and copy conditions of the Figure Copy Task. Impairment on the task was classified as greater than 2*SD*s below the mean score, and the overall graph takes into account both conditions. Figure available at https://osf.io/q6zys/ under a CC-BY4.0 license.

Table 3

Summary statistics of performance of the participants scored by human raters and the Automated Scoring Program,

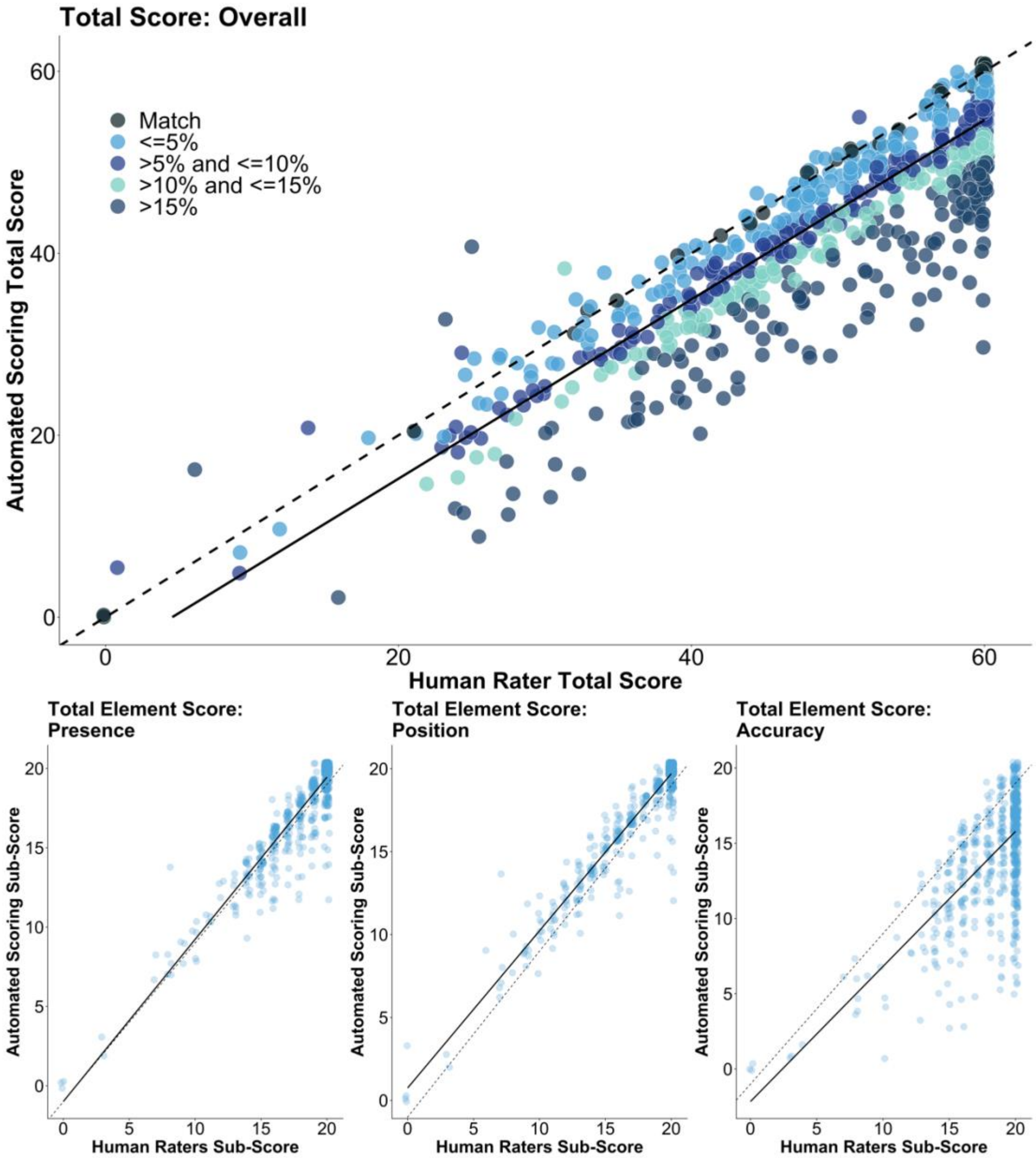
Condition	Group	Total Human score *M* (*SD*)	Algorithm score *M* (*SD*)	True positive	True negative	False positive	False negative	Sens.	Spec.
Figure Copy	Healthy adults	59.03 (2.90)	55.65 (4.68)	2.30	91.19	2.68	3.83	37.52	97.14
	Sub-Acute stroke	50.2 (13.45)	44.84 (11.62)	24.49	59.18	16.33	0	100	78.37
	Chronic Stroke	57.33 (5.98)	48.96 (7.95)	12.99	70.13	16.88	0	100	80.60
	Overall	57.72 (6.09)	52.29 (7.94)	8.19	80.82	8.84	2.16	79.13	90.14
Figure Recall	Healthy adults	47.83 (7.97)	43.95 (8.70)	1.53	95.40	1.15	1.92	44.35	98.81
	Sub-Acute stroke	32.36 (14.31)	31.57 (12.86)	26.53	71.43	2.04	0	100	97.22
	Chronic Stroke	41.69 (10.84)	34.13 (11.67)	18.83	71.43	5.84	3.90	82.84	92.44
	Overall	44.4 (10.69)	39.38 (11.49)	9.91	84.91	2.80	2.37	80.7	96.81

Note. We present sensitivity/specificity of the impairment classifications of the Automated Scoring Program (e.g., <2SDs from healthy adult mean total score) compared to ground truth impairment classifications of the manually scored total scores, per group and overall.

To further illustrate the degree of agreement between scoring methods overall in terms of total score we classified assigned automated scores into four categories: (1) a direct match with averaged rater total scores (2) within 5% of averaged rater total scores (3) between five (not inclusive) and 10% (inclusive) of averaged rater total scores, (4) between 10 and 15% deviation from averaged rater total scores, and (5) greater than 15% deviation from averages rater scores (e.g., Canham et al., 2000). We found that that 83.51% of scores from the algorithm were within 15% of the average rater scores (39.76% within 5%) and that the maximum deviation was 52% (*n*=1). In this single extreme case, the participant had drawn a non-element outside of the figure boundary, but within the maximum bounds, skewing the normalisation process such that the algorithm failed to recognise one side of the otherwise perfect figure. 16.16% were scored by the algorithm with a deviation greater than 15%. See Figure 5.

Figure 5

Illustrates the agreement between the human raters and the automated scoring program in both total scores and element-wise sub-scores



Note. The top panel illustrates the relationship between automated and average manual Figure Copy total scores (*N*=928). Lower panels present a comparison of accuracy, position, and presence element scores. The dashed line represents perfect agreement (slope =1, intercept = 0), and the linear best fit line is in black. This reveals the automatic scoring algorithm underscores, and that the underscoring is probably due to underscoring of accuracy. Figure available at https://osf.io/zy9ab/ under a CC-BY4.0 license.

Table 4 presents the automated program’s proportion of element hits, misses, false positives, and correct rejections for element-wise presence scores versus the human raters. Overall, the automated algorithm was found to exhibit an average element sensitivity of 90.10% and an average specificity of 92.20%. See supplementary materials for sensitivity tables for each element score (i.e., presence, accuracy, and position) and condition separately.

Table 4

The Automated Scoring Program's proportion of presence score hits, misses, false positives, and correct rejections versus the matched human rater data for combined copy and recall condition drawings

Element number	Element	True positive	True negative	False positive	False negative	Sens.	Spec.
1	Left vertical line	83.41	5.64	0.33	10.62	88.71	94.44
2	Left top horizontal line	87.25	3.96	0.33	8.46	91.16	92.31
3	Right top horizontal line	86.17	3.51	0.33	9.99	89.61	91.43
4	Right vertical line	85.35	5.73	0.33	8.59	90.86	94.55
5	Right bottom horizontal line	86.56	1.98	0.11	11.34	88.41	94.74
6	Left bottom horizontal line	87.05	2.31	0.22	10.43	89.30	91.30
7	Middle vertical line	95.48	1.87	0.33	2.31	97.64	85.00
8	Left bottom right slanted line	83.05	8.25	2.26	6.44	92.80	78.49
9	Left top right slanted line	80.92	11.83	1.23	6.03	93.07	90.60
10	Top left slanted line	71.48	21.25	2.57	4.70	93.83	89.20
11	Middle bottom right slanted line	78.07	13.89	2.18	5.86	93.02	86.43
12	Left small vertical line(1)	64.20	22.35	1.60	11.86	84.41	93.33
13	Left small vertical line(2)	64.74	23.58	1.81	9.86	86.78	92.86

Element number	Element	True positive	True negative	False positive	False negative	Sens.	Spec.
14	Circle	79.85	14.98	0.66	4.52	94.65	95.77
15	Right middle horizontal line(1)	90.81	4.48	1.12	3.59	96.20	80.00
16	Right middle horizontal line(2)	91.44	3.78	1.11	3.67	96.14	77.27
17	Star	85.48	11.64	0.11	2.77	96.86	99.06
18	Cross	82.96	13.92	1.78	1.34	98.41	88.65
19	Right bottom left slanted line	58.21	28.60	1.34	11.84	83.09	95.52
20	Right bottom half vertical line	78.80	17.54	1.22	2.44	96.99	93.49
	Average	81.06	11.05	1.05	6.83	92.10	90.20

Note. *Sens* = sensitivity, *Spec* = specificity.

Next, Cohen's *k* analyses were performed to evaluate the degree of agreement between automated and manual element accuracy, position, and presence scores. These scoring methods were found to exhibit a high degree of agreement on position (k=.82, 95% CI [.82-.84], p<.001) and presence (k=.76, 95% CI [.76-.77], p<.001) scores, but a lower degree of agreement within accuracy scores (k=.41, 95% CI [.41-.42], p<.001). The greatest source of disagreement between the automated and manual scorings was found to be element accuracy false positives (22.63% accuracy false positives versus 2.81% position and 4.72% presence false positives), resulting in a comparatively reduced overall specificity as seen in Table 4.

Algorithm versus nonmatched human data

Thus far, only data from element sub-scores in which human raters agreed with one another, or averaged total scores, has been considered. However, it is also important to investigate the performance of the automated program in more ambiguous cases. For this reason, the Automated Scoring Program was then evaluated within drawings for where human raters disagreed. To do this, we examined element or total score cases in which the two raters did not agree.

First, ICC analyses were conducted to identify the degree of agreement between automated scores and individual rater's assigned scores within cases where raters had assigned different total scores (37.07% of the time). Automated scores exhibited high consistency with both rater one's assigned scores (ICC = .80, *F*(553,553)=17.91, *p* <.001, 95% CI [.25-.92]) and rater two's (ICC = .79, *F*(553,553)=18.73, *p* <.001, 95% CI [.13-.92]) in cases where raters had assigned different total scores (*n* =344).

We then examined how the Automated Scoring Program resolved these rater disagreements. On cases where there was a clear disagreement between humans raters (i.e., three element points disagreed upon) or cases with clearer agreement between human raters (i.e., only one element point disagreed upon) the Automated Scoring Program tended to give more points. Cases in which raters disagreed on a single element sub-score were classed as one-point disagreements. Similarly, cases where raters disagreed on all three element sub-scores are termed three-point disagreements. Element 11 (middle bottom right slanted line) had the most three-point disagreements, and element 14 (the circle) had the most one-point disagreements. The algorithm scored the majority of participants three points (46.77% for element 11, 71.66% for element 14), and the consistency between the Automated Scoring Program and the average human rater score was moderate (element 11 ICC = .61, *F*(36,36)=8.32, *p* <.001, 95% CI [-.02-.85], element 14 ICC = .61, *F*(56,56)=10.44, *p* <.001, 95% CI [-.08- .85])).

However, in cases there the scoring was more ambiguous (i.e., disagreement by two-points was common on the small left vertical interior divider line, element 12), the automated scoring algorithm scored less favourable, giving the majority of participants zero points (37.39%). For these participants, however, the consistency between the Automated Scoring Program and the average human rater total score was still good (ICC = .69, *F*(40,40)=15.80, *p* <.001, 95% CI [-.07-.90]).

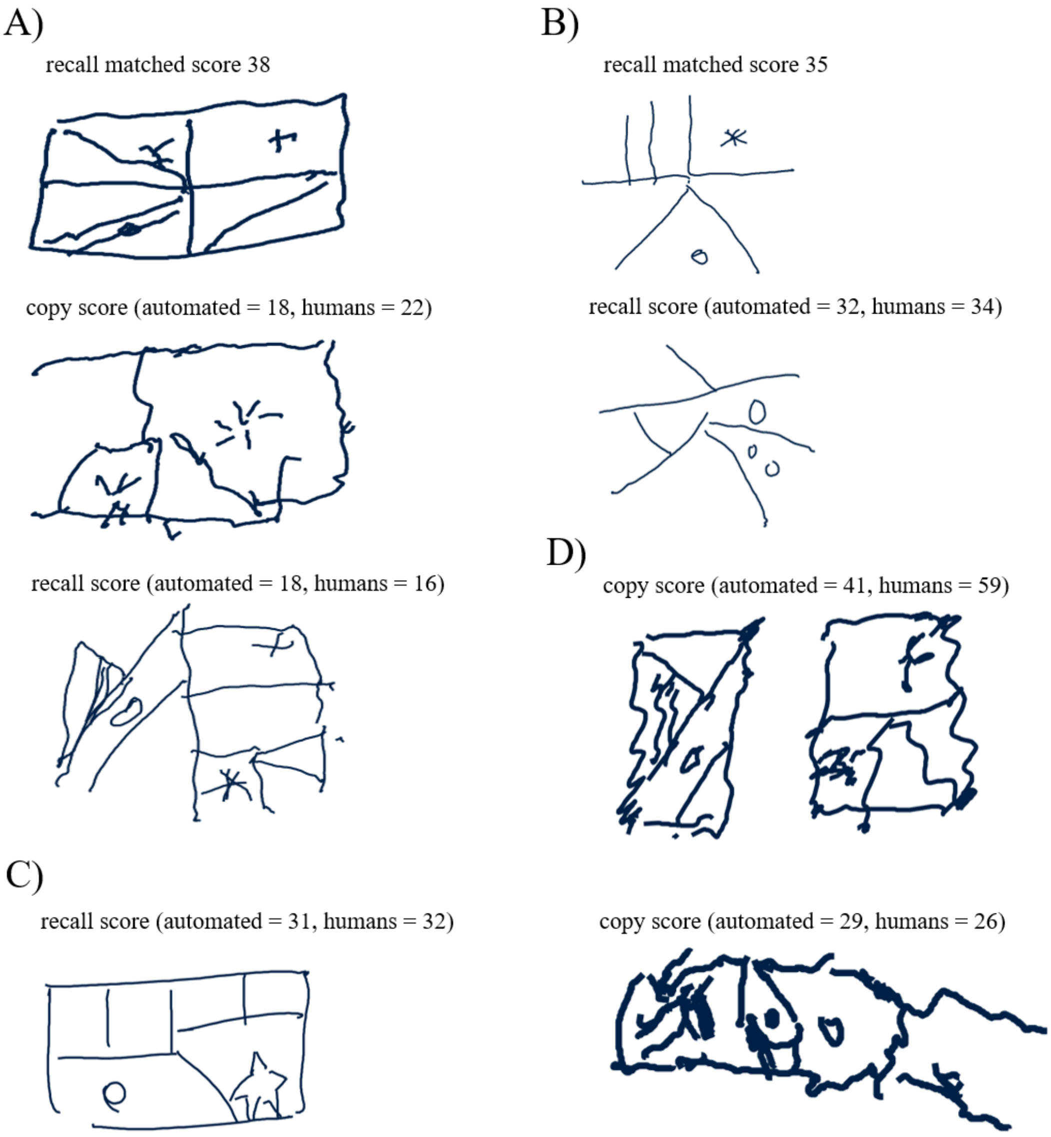
When looking at all elements score disagreements between raters regardless of the degree of disagreement (i.e., by one-point or more etc), the automatic scoring program proportionally gave less accuracy, position, and presence points (57.64% of the time) than awarded them (42.36% of the time), this was especially the case for accuracy where the Automated Scoring Program gave far more inaccuracy points than accuracy points. This is can be seen in Figure 4.

Strengths of the automated scoring program

Overall, the automated scorers matched well with the scores assigned by human raters with the majority of the automated total scores being within 15% of the manually assigned scores. The Automated Scoring Program was reliably able to extract and identify figure elements in drawings. For example, drawings which contained distorted or disconnected lines (Figure 6 panel A), partial copies (Figure 6 panel B), additional elements (Figure 6 panel C), and mild tremor (Figure 6 panel D) were generally scored accurately. Overall, the Automated Scoring Program was able to successfully discriminate elements from a wide range of imperfectly drawn figures.

Figure 6

Examples of distorted drawings from which the algorithm correctly identified and scored imperfectly drawn elements.


A)
recall matched score 38
copy score (automated = 18, humans = 22)
recall score (automated = 18, humans = 16)
B)
recall matched score 35
recall score (automated = 32, humans = 34)
C)
recall score (automated = 31, humans = 32)
D)
copy score (automated = 41, humans = 59)
copy score (automated = 29, humans = 26)

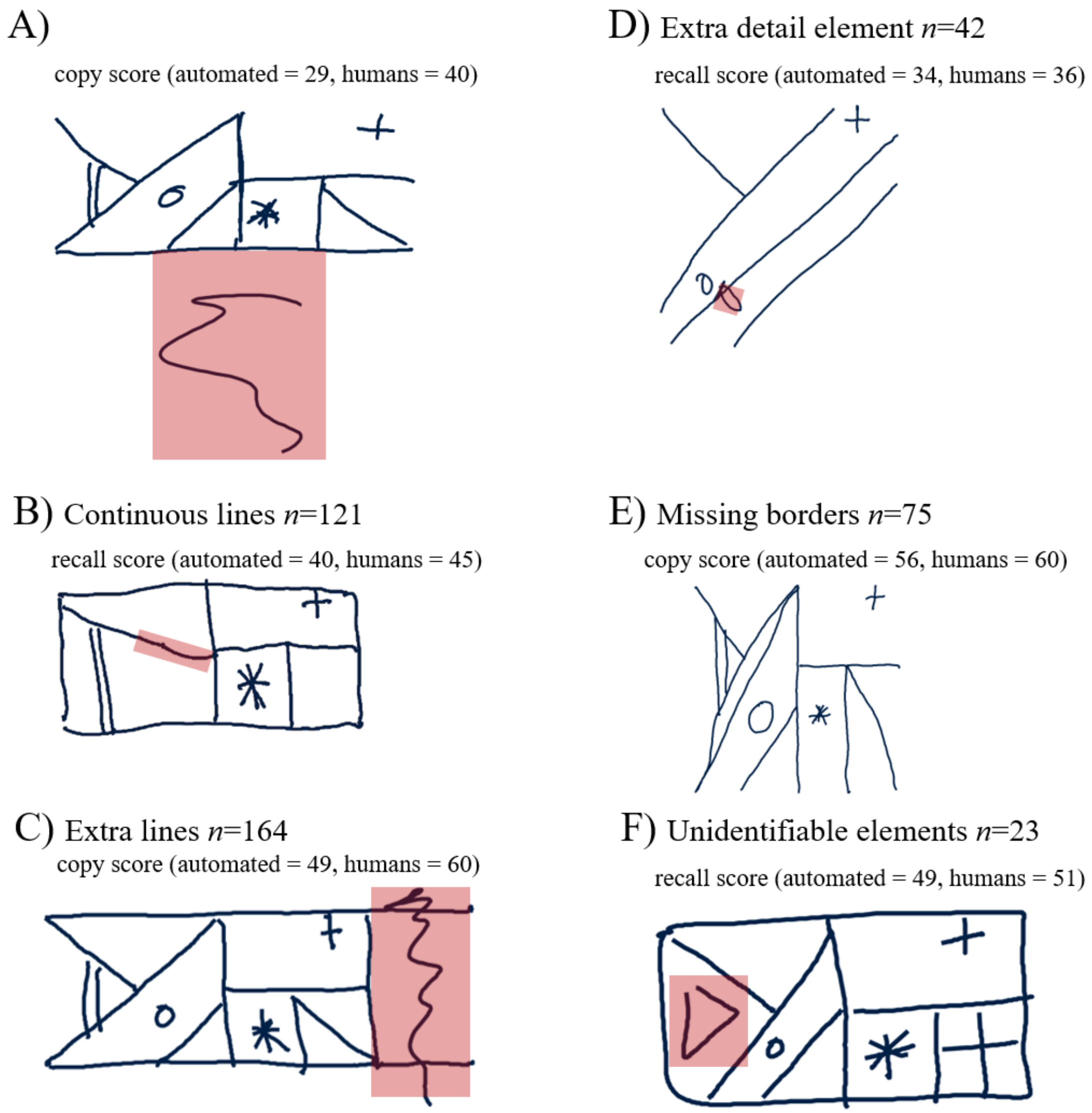
Note. Figure available at https://osf.io/7ks3g/ under a CC-BY4.0 license.

Automated Scoring Challenges

The ability of the Automated Scoring Program to effectively quantify Figure Copy Task performance can be additionally investigated by determining specific cases in which this algorithm struggles to accurately identify all figure elements. Although this program is generally robust with <10.06% element false positives and <1.17% element false negatives, there are specific response patterns which may result in systematic scoring failures. For example, the scoring algorithm is less accurate in cases where extra elements were included (Figure 7 Panels A, D, F). The Automated Scoring Program also struggles in cases where participants extend line elements beyond their template boundaries (Figure 7, Panel B), had attempted to correct mistakes (Figure 7, Panel C), or had failed to draw figure border elements (Figure 7, Panel E).

Figure 7

Sample figures in which response features caused inaccuracies within the Automated Scoring Program.

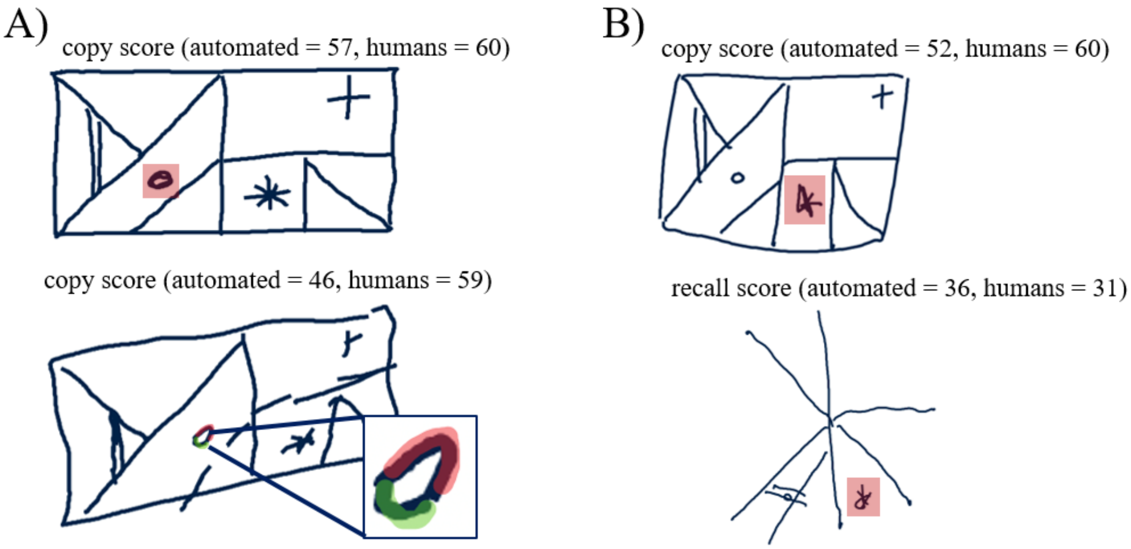

A)
copy score (automated = 29, humans = 40)
D) Extra detail element *n*=42
recall score (automated = 34, humans = 36)
B) Continuous lines *n*=121
recall score (automated = 40, humans = 45)
E) Missing borders *n*=75
copy score (automated = 56, humans = 60)
C) Extra lines *n*=164
copy score (automated = 49, humans = 60)
F) Unidentifiable elements *n*=23
recall score (automated = 49, humans = 51)

Note. Elements which distorted the automated scoring process are highlighted in red. The Automated Scoring Program was found to struggle when additional figure elements were drawn (A, C, D, F), when borders were missing (E), or when elements were extended beyond their template boundaries (B). Number of participants who the raters identified as making these errors are noted in panel labels. Note panel E shows a drawing with a perfect score, due to awarding full border element points if no border element is present and/or the participant used the edge of the on-screen drawing area as the figure border. Figure available at https://osf.io/8ahtu/ under a CC-BY4.0 license.

The Automated Scoring Program was found to encounter the most difficulty when scoring circle and star elements, as the algorithm must employ precise criteria (e.g., number and angle of intersecting lines) to identify these features. For example, the automated algorithm struggled to identify the circular element, missing 7.22% of circles which were marked as present by human raters. This systematic false negative specifically occurred when the circle was drawn as an arc, as multiple distinct overlapping lines, or another non-closed path (Figure 8 panel A). Star elements also may not be correctly identified, with the automated program missing 5.60% of stars marked as present by human rater. This false negative occurs if stars are drawn as a single continuous path, rather than as distinct lines (Figure 8 panel B). However, it should be noted that overall the inaccurate scoring by the algorithm was comparatively infrequent, impacting scoring on only 119/928 (12.82%) drawings considered.

Figure 8

Further examples in which response features caused inaccuracies within the Automated Scoring Program.


A)
copy score (automated = 57, humans = 60)
copy score (automated = 46, humans = 59)
B)
copy score (automated = 52, humans = 60)
recall score (automated = 36, humans = 31)

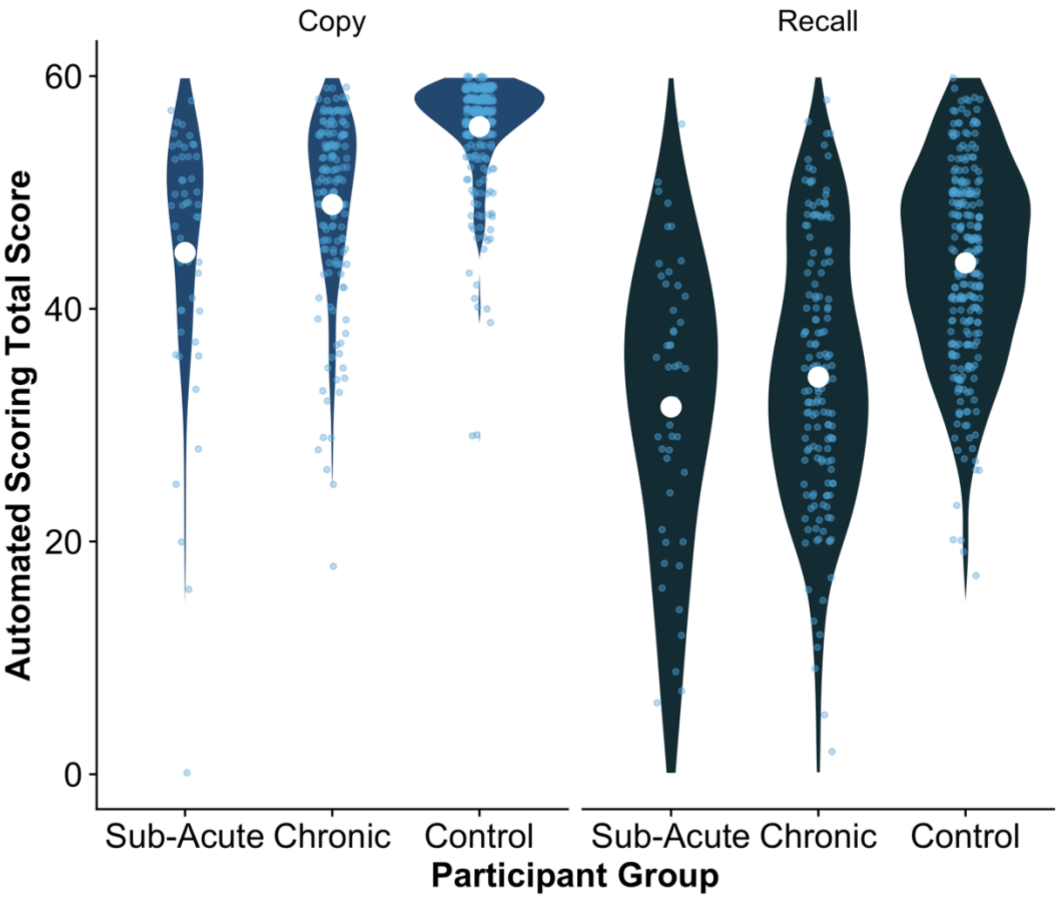
Note. Illustrates examples of figures where circular elements (Panel A) and star elements (Panel B) were not identified by the Automated Scoring Program. Circles which were drawn as multiple, overlapping lines or multiple separated lines were not identified (Panel A, distinct lines highlighted in red and green, or for black and white print, in disjointed translucent highlight). Similarly, stars which were drawn as a continuous line were missed by the Automated Scoring Program (Panel B). Figure available at https://osf.io/8b2dn/ under a CC-BY4.0 license.

Known-group discriminability

In order to sanity check the scoring of the Automated Scoring Program, we compared the three sample groups (i.e., sub-acute stroke, chronic stroke, and healthy adults), to see if they performed differently from each other. Given typical recovery trajectories following stroke, we would expect that the sub-acute stroke group would score lower than the chronic stroke group, and that the chronic stroke group would score lower than the healthy adult group. An ANCOVA analysis was conducted to establish the differences between the healthy adult and stroke survivor groups in their total scores while controlling for statistical differences of demographics of age and education. For the copy condition, the ANCOVA revealed a significant effect of group, when controlling for age and education (*F*(2,386)=40.80, *p* <.001). Tukey HSD test indicated that healthy adults performed significantly better than both sub-acute stroke survivors ($M_{\text{difference}}$=7.65, *p*<.001, *d*=-1.72) and chronic stroke survivors ($M_{\text{difference}}$=3.96, *p*<.001, *d*=-1.10). For stroke survivors specifically, lesion volume was added to the model as a covariate and Tukey HSD demonstrated significant differences between the sub-acute and chronic stroke survivors (M $_{\text{difference}}$=4.69, *p*=.01, *d*=-.46). For the recall condition, when only controlling for the effects of age and education on total score, healthy adults again performed significantly better in the recall condition than both sub-acute stroke survivors ($M_{\text{difference}}$=8.94, *p*<.001, *d*=-1.31) and chronic stroke survivors ($M_{\text{difference}}$=4.63, *p*<.001, *d*=-.99). When additionally adding lesion volume, on this recall data, the Tukey HSD analysis revealed no significant difference in performance on the recall conditions ($M_{\text{difference}}$=3.89, *p*=.13, *d*=-.21). See Figure 9 for the distributions of total score on the copy and recall Figure Copy test conditions.

Figure 9

Illustrates the distributions of total score from three groups on the Figure Copy test, both copy and recall conditions.



Note. White dot represents separate group means, individual scores are in light blue, and score distributions are presented in darker blue. Sub-Acute and Chronic refer to stroke survivors before and after 2 months post-stroke. Figure available at https://osf.io/pfmcu/ under a CC-BY4.0 license.

Discussion

This investigation aimed to develop a novel, automated program to score the OCS-Plus Figure Copy Task (Demeyere et al., in press) and to evaluate the accuracy and utility of this automated tool. Overall, the automated scoring algorithm was able to reliably extract and identify individual figure elements and to assign total scores which agreed well with manual scores across both the copy and recall conditions. Compared to overall impairment categorisations based on manual scores, the Automated Scoring Program had a high overall sensitivity and specificity and was reliably able to distinguish between different clinical impairment groups. The novel automated program was found to be generally robust and very close to the manual scoring overall. There is a clear benefit of automating Figure Copy scoring, in terms of time and cost savings, in particular allowing this screening assessment to be used without the need for highly trained neuropsychologists to administer and score the task. At the group level, the scoring tool is clearly able to distinguish groups, and diagnostic accuracy compared to manual scoring was very high with an overall AUC of 86.7%. At an individual patient level, we did note some specific response patterns which resulted in systematic scoring failures on the automatic tool. Even though these were low in incidence, if the scoring algorithm is to be used at an individual diagnostic level, it is important to highlight these. Combining the automated scores with full visualization of the original drawing in the reports is key to help interpret all scores at the individual level (see https://osf.io/y9xvk/ for an example output of the Automated Scoring Program). Overall, the very high alignment with manual scoring means this program represents a significant and pragmatic advancement over traditionally employed manual scoring procedures, setting the scene for potential implementation in wide-scale screening programmes, potentially even in self-assessment settings.

Within this investigation, the two human raters were found to assign scores with a high degree of agreement. This consistency was present across individual element sub-scores as well as within both copy and recall condition data. When human raters did not assign identical scores, the source of disagreement was most commonly individual element position and accuracy scores. However, the human raters in this investigation completed an extensive training program designed to standardize assigned scores which is typically not feasible to implement at scale within clinical environments. In order to allow a more automated and wide scale range of cognitive screening to be conducted, reducing the need for high level training and neuropsychologists to score such tasks, is a pragmatic solution.

Overall, the automated scoring algorithm was able to reliably extract and accurately score individual elements within patient Figure Copy Task responses. In cases where human raters assigned identical scores, there was a high degree of consistency between automated and manually assigned total scores and moderate agreement within individual element scores. The overall human-algorithm score correlations in this investigation were largely similar to those reported by Vogt et al. (2019) (.83 versus .88 respectively). Within individual elements, the Automated Scoring Program demonstrated extremely high sensitivity (92.10%) and specificity (90.20%). It is important to note that human-algorithm scoring differences do not necessarily represent algorithmic errors, but instead suggest the use of slightly different, but not necessarily less valid, scoring criteria. For example, the automated program has a tendency to be stricter than humans raters when awarding points within the accuracy element sub-score. However, despite this systematic difference within accuracy sub-scores, the vast majority of automated total scores (83.51%) were within 15% of manually assigned total scores. These findings suggest that the automated program employs slightly different element scoring criteria than the human raters, but this variance does not result in substantial changes within response total scores.

The performance of the automated scoring algorithm was also separately investigated within responses where human raters did not assign identical scores. This is a particularly critical analysis to conduct, as an automated algorithm can potentially provide a systematic, reproducible method for resolving such human rater disagreements. Within individual response elements which were assigned different scores by human raters, the automated program tended to employ more lenient scoring criteria. For example, when a specific element was scored as being inaccurately drawn by one rater but accurately drawn by the other, the Automated Scoring Program was more likely to report that the element had been drawn accurately. Despite this tendency to be more lenient, as a whole, automated scores exhibited high consistency with both rater one and rater two's assigned total scores in cases where both raters disagreed. This indicates that the automated program's systematic lenient scoring of disagreed upon individual elements does not appear to produce systematic biases within overall response scores. In any case where quantitative scores are assigned to response which do not have a clear "ground truth" score, some degree of subjective judgement is required. The Automated Scoring Program employs consistency where human raters may not and provides the clear advantage of being able to standardize scoring across all responses.

The Automated Scoring Program was found to exhibit several clear strengths over manual scoring procedures. First, was able to systematically assign completely reproducible scores even in cases where drawings were distorted. Given that this investigation included data from a representative sample of sub-acute stroke survivors exhibiting a range of common post-stroke cognitive impairments, responses were frequently extremely dissimilar to the target figure.

The automated program was found to cope well with drawing inaccuracies due to co-morbid fine motor impairments, omissions due to visuospatial deficits, perseveration errors, and other common post-stroke impairment patterns. This robustness greatly adds to the automated program's potential clinical utility. Second, while manual scoring of Complex Figure Copy drawing requires training and time to complete, the automated program is able to instantly produce detailed score breakdowns. This makes employing an automated scoring procedure extremely time efficient, which is a valued attribute especially within clinical settings. Finally, the scores generated by the automated algorithm are completely reproducible. These standardized scores are one of the greatest advantages of employing automated over manual scoring methods, as they facilitate valid score comparisons across many different raters in many different patient groups.

Despite these advantages, some potential weaknesses were identified within the automated scoring procedure. First, there are specific response patterns which were found to result in systematic underscoring. For example, the automated program struggled to identify circle and star elements which did not meet its exact mathematical extraction criteria but were easily identifiable by human raters (Figure 8). Similarly, the Automated Scoring Program struggled to accurately score drawings when large, extra features were present within the response space (Figure 7). These failures occurred infrequently but represent a potential avenue for improving the automated scoring procedure, or even simply providing an extra element of confidence ratings for each figure, to flag up those which may have been underscored. Future research should aim to identify more flexible methods for identifying more complex elements and for preventing the presence of large extra elements from distorting figure segmentation. Finally, the automated program employs slightly different element sub-scoring strategies than human raters. Where the circle, star, and cross elements can be identified by the automated program, it is automatically scored as “accurate” due to the equations having specific placement and line intersection requirements. This means that for these specific 3 detail elements, they cannot be scored as present correctly if they are not also scored as being accurate. However, this difference in scoring was not found to result in significant disagreements between automated and human-assigned scores.

Several previous investigations of automated segmentation algorithms have found that the best results are achieved when scoring procedures employ limited human feedback to address minor weaknesses in otherwise robust algorithms. For example, Wang et al (2016) developed a deep learning algorithm to identify and segment potentially cancerous tissue in mammograms which found that a trained pathologist achieved an AUC of 99.6% whilst the automated segmentation program achieved and AUC of 96.6%. However, when the automated output was briefly reviewed by the trained pathologist to remove obvious false positive cell clusters, the maximal AUC of 99.5% was achieved whilst retaining the time-efficiently benefit of employing an automated scoring method (Wang et al., 2016). A similar approach could potentially be taken to improve the performance of the automated scoring program presented within this investigation. For example, human raters could quickly screen all figures assigned very low scores by the algorithm to flag cases where normalizing errors have produced false negative scores. Overall, the automated scoring program was found to provide a robust and reliable method for analyzing a wide range of Figure Copy Task responses. However, future investigations can aim to further explore clinical feasibility and acceptability and within this investigate whether employing a collaborative scoring approach could maximize the efficacy and accuracy of automated scoring processes. Importantly, the automated scores were found to reliably distinguish between participants falling into different impairment groups. On average, sub-acute stroke survivors were assigned significantly lower scores than chronic stroke survivors, who were in turn assigned lower average scores than neurologically healthy adults. These findings are in line with expectations, demonstrating the external validity of automated Figure Copy Task Scores. Receiver Operating Curve analysis demonstrated that, compared to overall impairment categorizations based on manual scores, the Automated Scoring Program had an extremely high overall sensitivity and specificity (80% and 93.4% respectively; AUC = 86.7%). This finding illustrates that impairment classifications based on automated scores alone are largely comparable to those assigned by human raters. Taken together, this external validity and ability to identify overall impairment highlight the automated scorning program’s potential clinical utility.

Complex Figure Copy Tasks are commonly used as a component of neuropsychological evaluations within both clinical and research settings. From a clinical perspective, automated scoring offers a time-efficient solution for standardizing Figure Copy Scores in order to more reliably detect impairment patterns across many different patient groups. Examiners will no longer have to complete time-consuming scoring or training procedures, and will be provided with immediate, highly detailed scoring results. This in turn may help improve the speed and accuracy of identifying common visuospatial and non-verbal memory based neuropsychological deficits and open the door to wider population-based cognitive screening and (assisted) self-assessments. From a research perspective, employing automated Figure Copy Scoring helps reduce bias due to the reliance on subjective examiner judgments. This is a critical advantage, as it facilitates valid, large-scale comparisons of Figure Copy Task data collected by different examiners, within different patient groups or research settings. Automated scoring is also completely reproducible, augmenting the reliability of any findings based on analysis this scoring data. Overall, the results of this investigation strongly suggest that the novel, automated Figure Copy Scoring tool is a robust and reliable scoring methodology which can be employed to produce immediate, unbiased, and reproducible scores for Complex Figure Copy Task responses in clinical and research environments.

Limitations

There are several potential avenues through which future research can aim to expand on the findings of this investigation. First, Complex Figure Copy Tasks are not only commonly employed within stroke patients, are also regularly administered to patients with suspected dementia, traumatic brain injury, and other neurological deficits. Patients falling within each of these impairment categories may exhibit different error patterns within Figure Copy Tasks. Future research can aim to investigate whether this Automated Scoring Program performs equally well across these patient groups and to determine whether these Figure Copy tasks can reliably differentiate between a wider range of clinical populations. Second, this Automated Scoring Program was specifically designed to score the computerised OCS-Plus Figure Copy Task and does not generalise to other Figure Copy Task Stimuli. In order to facilitate automated scoring of other common Figure Copy Tasks (such as the Rey-Ossterich Complex Figure), future research will need to develop additional, specialised automated scoring algorithms. Similarly, the automated algorithm relies on detailed (x,y) coordinates and timestamps produced by a tablet computer-based task. While computerised neuropsychological testing is rapidly being adapted in clinical and research environments (e.g., Bauer et al 2012), many Figure Copy Tasks are still administered in pen and paper format, and the embedding of computerised testing in clinical practice remains a challenge (e.g., see Schmand, 2019).

Conclusions

This investigation presents a novel, automated scoring tool for the OCS-Plus Figure Copy Task (Demeyere et al., in press). Overall, the automated scoring algorithm was able to reliably extract and identify individual figure elements and to assign total scores which agreed well with manual scores across both the copy and recall conditions. This automated program was reliably able to identify overall impairment patterns and distinguish between different clinical impairment groups. This represents a significant advancement as this novel technology can be employed to produce immediate, unbiased, and reproducible scores for Complex Figure Copy Task responses in clinical and research environments. More generally, the findings of this investigation suggest that automated scoring procedures can be implemented to improve the scope and quality of neuropsychological investigations by reducing reliance on subjective examiner judgments and improving scoring time-efficiency.

References

Agrell, B., & Dehlin, O. (1998). The clock-drawing test. *Age and Ageing, 27*(3), 399–403.

Alladi, S., Arnold, R., Mitchell, J., Nestor, P. J., & Hodges, J. R. (2006). Mild cognitive impairment: Applicability of research criteria in a memory clinic and characterization of cognitive profile. *Psychological Medicine*, *36*(4), 507–515. https://doi.org/10.1017/S0033291705006744

Al-Janabi, H., Flynn, T. N., & Coast, J. (2012). Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. *Quality of life research*, *21*(1), 167-176.

Barker, L. A., Morton, N., Morrison, T. G., & McGuire, B. E. (2011). Inter-rater reliability of the Dysexecutive Questionnaire (DEX): Comparative data from non-clinician respondents—all raters are not equal. *Brain Injury, 25*(10), 997–1004. https://doi.org/10.3109/02699052.2011.597046

Canham, R., Smith, S. L., & Tyrrell, A. M. (2000). Automated scoring of a neuropsychological test: The Rey Osterrieth complex figure. *Proceedings of the 26th Euromicro Conference. EUROMICRO 2000. Informatics: Inventing the Future*, *2*, 406–413 vol.2. https://doi.org/10.1109/EURMIC.2000.874519

Chechlacz, M., Novick, A., Rotshtein, P., Bickerton, W.-L., Humphreys, G. W., & Demeyere, N. (2014). The neural substrates of drawing: A voxel-based morphometry analysis of constructional, hierarchical, and spatial representation deficits. *Journal of Cognitive Neuroscience*, *26*(12), 2701–2715. https://doi.org/10.1162/jocn_a_00664

Chen, H., Pan, X., Lau, J. K. L., Bickerton, W.-L., Pradeep, B., Taheri, M., Humphreys, G., & Rotshtein, P. (2016). Lesion-symptom mapping of a complex figure copy task: A large-scale PCA study of the BCoS trial. *NeuroImage. Clinical*, *11*, 622–634. https://doi.org/10.1016/j.nicl.2016.04.007

Chen, S., Stromer, D., Alabdalrahim, H. A., Schwab, S., Weih, M., & Maier, A. (2020). Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports*, *10*(1), 20854. https://doi.org/10.1038/s41598-020-74710-9

Crawford, J. R., Parker, D. M., McKinnley, W., & McKinlay, W. W. (1992). *A Handbook of Neuropsychological Assessment*. Psychology Press.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674.

Demeyere, N., Haupt, M., Webb, S. S., Strobel, L., Milosevich, E. T., Moore, M. J., Wright, H., Finke., Duta, M. D. (in press). Introducing the tablet-based Oxford Cognitive Screen-Plus (OCS-Plus) as an assessment tool for subtle cognitive impairments. *Scientific Reports.*

Demeyere, N., Riddoch, M. J., Slavkova, E. D., Bickerton, W. L., & Humphreys, G. W. (2015). The Oxford Cognitive Screen (OCS): validation of a stroke-specific short cognitive screening tool. *Psychological assessment*, *27*(3), 883.

Deza, M. M., & Deza, E. (2009). Encyclopedia of distances. In *Encyclopedia of distances* (pp. 1–583). Springer.

Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *10*(2), 112–122.

Duncan, P. W., Reker, D. M., Horner, R. D., Samsa, G. P., Hoenig, H., LaClair, B. J., & Dudley, T. K. (2002). Performance of a mail-administered version of a stroke-speci” c outcome measure, the Stroke Impact Scale. *Clinical rehabilitation, 16*(5), 493-505.

Ellis, A. W., & Young, A. W. (2013). *Human Cognitive Neuropsychology: A Textbook With Readings*. Psychology Press.

Fennema, J. (n.d.). PyLaTeX. *Https://Github.Com/JelteF/PyLaTeX*.

Folstein, M. F., Robins, L. N., & Helzer, J. E. (1983). The mini-mental state examination. *Archives of General Psychiatry*, *40*(7), 812–812.

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association : JMLA*, *105*(2), 203–206. https://doi.org/10.5195/jmla.2017.88

Franzen, M. D. (2000). *Reliability and Validity in Neuropsychological Assessment*. Springer Science & Business Media.

Freeman, R. Q., Giovannetti, T., Lamar, M., Cloud, B. S., Stern, R. A., Kaplan, E., & Libon, D. J. (2000). Visuoconstructional problems in dementia: Contribution of executive systems functions. *Neuropsychology*, *14*(3), 415.

Gamer, M., Lemon, J., Fellows, I., & Puspendra, S. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (0.84.1) [R]. https://CRAN.R-project.org/package=irr

Gillies, S. (2015). The Shapely User Manual Shapely 1.2 and 1.3 documentation. *Zugriff Am*, 11–16.

Humphreys, G. W., Duta, M. D., Montana, L., Demeyere, N., McCrory, C., Rohr, J., Kahn, K., Tollman, S., & Berkman, L. (2017). Cognitive function in low-income and low-literacy settings: Validation of the tablet-based Oxford Cognitive Screen in the Health and Aging in Africa: A Longitudinal Study of an INDEPTH community in South Africa (HAALSI). *The Journals of Gerontology: Series B*, *72*(1), 38–50.

Huygelier, H., Moore, M. J., Demeyere, N., & Gillebert, C. R. (2020). Non-spatial impairments affect false-positive neglect diagnosis based on cancellation tasks. *Journal of the International Neuropsychological Society*.

Jones, E., Oliphant, P., & Peterson, P. (2001). SciPy: Open source scientific tools for Python. *Www.Scipy.Org*.

Kassambara, A. (2020). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests: Vol. 0.4.0*.

Koppitz, E. M. (1964). *The Bender Gestalt test for young children.*

Mahoney, F., & Barthel, D. W. (1965). Functional evaluation ; the Barthel index. A simple index of the independence useful in scoring improvement in the rehabilitation of the chronically ill. *Maryland State Medical Journal, 14*, 61–65.

Maneewongvatana, S., & Mount, D. M. (2002). Analysis of approximate nearest neighbor searching with clustered point sets. *Data Structures, Near Neighbor Searches, and Methodology, 59*, 105–123.

Mangiafico, S. (2019). *rcompanion: Functions to Support Extension Education Program Evaluation* (2.3.7) [R]. https://CRAN.R-project.org/package=rcompanion

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press. https://doi.org/10.1037/11281-000

Moetesum, M., Siddiqi, I., Masroor, U., & Djeddi, C. (2015). Automated scoring of Bender Gestalt Test using image analysis techniques. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 666–670. https://doi.org/10.1109/ICDAR.2015.7333845

Moore, M. J., Vancleef, K., Shalev, N., Husain, M., & Demeyere, N. (2019). When neglect is neglected: NIHSS observational measure lacks sensitivity in identifying post-stroke unilateral neglect. *J Neurol Neurosurg Psychiatry*, jnnp-2018-319668. https://doi.org/10.1136/jnnp-2018-319668

Pinto, E., & Peters, R. (2009). Literature review of the Clock Drawing Test as a tool for cognitive screening. *Dementia and Geriatric Cognitive Disorders*, *27*(3), 201–213.

Power, M., Fell, G., & Wright, M. (2013). Principles for high-quality, high-value testing. *BMJ Evidence-Based Medicine*, *18*(1), 5–10.

R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing,. https://www.R-project.org/

Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in Test-Usage Practices of Clinical Neuropsychologists in the United States and Canada Over a 10-Year Period: A Follow-Up Survey of INS and NAN Members. *Archives of Clinical Neuropsychology*, *31*(3), 206–230. https://doi.org/10.1093/arclin/acw007

Revelle, W. (2018). *psych: Procedures for personality and psychological research. Northwestern University, Evanston*.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 1–8.

Schreiber, H. E., Javorsky, D. J., Robinson, J. E., & Stern, R. A. (1999). Rey-Osterrieth Complex Figure Performance in Adults with Attention Deficit Hyperactivity Disorder: A Validation Study of the Boston Qualitative Scoring System. *The Clinical Neuropsychologist*, *13*(4), 509–520. https://doi.org/10.1076/1385-4046(199911)13:04;1-Y;FT509

Shin, M.-S., Park, S.-Y., Park, S.-R., Seol, S.-H., & Kwon, J. S. (2006). Clinical and empirical applications of the Rey–Osterrieth Complex Figure Test. *Nature Protocols*, *1*(2), 892–899. https://doi.org/10.1038/nprot.2006.115

Somerville, J., Tremont, G., & Stern, R. A. (2000). The Boston qualitative scoring system as a measure of executive functioning in Rey-Osterrieth complex figure performance. *Journal of Clinical and Experimental Neuropsychology*, *22*(5), 613–621.

Taylor, L. B. (1969). Localisation of cerebral lesions by psychological testing. *Clinical Neurosurgery*, *16*, 269–287.

Virbel, M., Hansen, T., & Lobunets, O. (2011). Kivy–a framework for rapid creation of innovative user interfaces. *Workshop-Proceedings Der Tagung Mensch & Computer 2011. ÜberMEDIEN| ÜBERmorgen*.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *ArXiv Preprint ArXiv:1606.05718*.

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist*, *31*(6–7), 1113–1126. https://doi.org/10.1080/13854046.2017.1317364

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. springer.

Wilke, C. O. (2019). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2": Vol. 1.0.0*.

Yamshchikova, A. (2019). *Automated scoring: An investigation into Figure Copy Task, a neuropsychological drawing test* [Masters Dissertation]. Anonymous university

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta psychiatrica scandinavica*, *67*(6), 361-370.