

# SCIENTIFIC REPORTS



OPEN

## Picture Interpretation Test (PIT) 360°: An Innovative Measure of Executive Functions

Silvia Serino<sup>1,2</sup>, Francesca Baglio<sup>3</sup>, Federica Rossetto<sup>2,3</sup>, Olivia Realdon<sup>4</sup>, Pietro Cipresso<sup>1,2</sup>, Thomas D. Parsons<sup>5,6</sup>, Giacomo Cappellini<sup>7</sup>, Fabrizia Mantovani<sup>4</sup>, Gianluca De Leo<sup>8</sup>, Raffaello Nemni<sup>3,9</sup> & Giuseppa Riva<sup>1,2</sup>

The assessment of executive functions poses researchers with several challenges related to both the complexity of the construct of executive functions itself and/or the methodological difficulties related to its evaluation. The main objective of the current study was to evaluate a 360° version of an ecologically valid assessment called the Picture Interpretation Test (PIT). Participants included 19 patients with Parkinson's disease (PD) and 19 healthy controls. All participants endorsed globally positive experiences of the PIT 360°. Furthermore, findings indicated that patients with PD took longer to correctly interpret the PIT 360° scene and tended to significantly focus on details of the 360° scene instead of the most informative elements. The time needed for a correct interpretation of the presented scene also correlated significantly with performance in conventional paper and pencil tests of executive functions for patients with PD. Classification analysis indicated the potential of the PIT 360° for distinguishing between patients with PD and healthy controls. Overall, these data provide preliminary evidence in support of the PIT 360° for evaluating executive functions.

The assessment of executive functions poses researchers with several challenges related to both the complexity of the construct of executive functions itself (see for example<sup>1</sup>) and/or the methodological difficulties related to its evaluation, specifically in predicting behaviors in real-life contexts<sup>2-9</sup>. In term of complexity, Chan and co-workers<sup>10</sup> have described executive functions as “an umbrella term comprising a wide range of cognitive processes and behavioral competencies which include verbal reasoning, problem-solving, planning, sequencing, the ability to sustain attention, resistance to interference, utilization of feedback, multitasking, cognitive flexibility, and the ability to deal with novelty” (pg. 201). Others have proposed four cognitive constructs: volition, planning, purposive action, and effective performance<sup>11</sup>. Still others reduce this list to three cognitive constructs, namely inhibitory control, working memory, and cognitive flexibility<sup>12-14</sup>.

In an attempt to refine executive functions assessment, Burgess *et al.*<sup>8</sup> advanced the idea of developing neuropsychological assessments based on models derived from directly observable everyday behaviors. Such an approach allows for an examination of the ways in which a sequence of actions leads to a given behavior in normal functioning. This “function-led” approach differs from the emphasis on abstract cognitive “constructs” without regard for their ability to predict the complexity of “functional” behaviors found in real-life situations<sup>2-5</sup>. Burgess and colleagues advanced the Multiple Errands test as a measure of executive functioning in real-life scenarios. While there are notable aspects of such naturalistic assessments, there are concerns about their many limitations in terms of time consumption, cost, poor control, and lack of safety<sup>15</sup>.

<sup>1</sup>Applied Technology for Neuro-Psychology Lab, IRCCS Istituto Auxologico Italiano, Via Magnasco, 2 20149, Milan, Italy. <sup>2</sup>Department of Psychology, Università Cattolica del Sacro Cuore, Largo Gemelli, 1, 20100, Milan, Italy. <sup>3</sup>IRCCS, Fondazione don Carlo Gnocchi ONLUS, Via Capecelatro 66, 20148, Milan, Italy. <sup>4</sup>Department of Human Sciences for Education, Università degli Studi di Milano-Bicocca, Milan, Italy. <sup>5</sup>Computational Neuropsychology and Simulation Laboratory, University of North Texas, 1155 Union Circle #311280, Denton, Texas, 76203-5017, USA. <sup>6</sup>Department of Psychology, University of North Texas, 1155 Union Circle #311280, Denton, Texas, 76203-5017, USA. <sup>7</sup>National Research Council of Italy, Institute for the Dynamics of Environmental Processes, Piazza della Scienza, 1, 20126, Milan, Italy. <sup>8</sup>Department of Clinical and Digital Health Sciences, College of Allied Health Sciences, Augusta University, 987 St. Sebastian Way, EC 4316, Augusta, Georgia, 30912, USA. <sup>9</sup>Department of Pathophysiology and Transplantation, Università degli Studi di Milano, via Francesco Sforza, 35, 20122, Milan, Italy. Correspondence and requests for materials should be addressed to S.S. (email: [s.serino@auxologico.it](mailto:s.serino@auxologico.it))

Received: 22 March 2017

Accepted: 1 November 2017

Published online: 22 November 2017

A potential alternative is the use of virtual reality (VR) technology for function-led assessments of executive functioning<sup>16</sup>. Indeed, VR permits the development of such assessments simulating everyday activities, allowing a secure and ecologically valid measure of executive functions<sup>9,17</sup>. A virtual Multiple Errands Test (VMET) has been developed and tested in various clinical populations<sup>18–20</sup>. The VMET allows for the evaluation of patients' abilities in formulating and checking a list of goals to effectively respond to environmental demands to achieve a series of tasks (e.g., buy a specific product, ask the examiner information about a product to be purchased). A recent study found that VMET was an effective tool for detecting early executive deficits in non-demented patients with Parkinson's Disease (PD)<sup>19</sup>. Furthermore, results demonstrated that patients with PD made more errors in the VMET tasks and showed a poorer ability in using effective strategies to complete the tasks in comparison to a control group. Interestingly, these two groups did not differ in their performance when compared on a traditional assessment of executive functioning.

A new technology for presenting neuropsychological stimuli is found in 360° environments (immersive photographs or videos) delivered via smartphones. The potentiality of 360° technology can be better understood by considering the “virtuality continuum” proposed by Milgram<sup>21</sup>, in which stimuli are presented in a manner ranging from completely real (real environment) to “virtual” (virtual environment). The space between extremes, called “mixed reality”, is the area wherein real and virtual may co-exist producing new experiences. Advances in 360° technologies allow participants to be immersed into a real situation that they experience from a first-person perspective. This platform allows for sequential focusing upon various elements and portions of the environment at different times. Moreover, this permits a sequential planning of visual search.

In this direction, we developed a 360° version of the Picture Interpretation Test (PIT)<sup>22,23</sup> that leverages Luria and colleagues<sup>24</sup> use of a Russian picture entitled “Unexpected Return” to investigate active visual perception in patients with frontal lobe damage. The approach was developed from their belief that the interpretation of a meaningful picture engages the patient's neurocognitive system via recursive selection of the most informative elements observed during the visual search. This allows for the elaboration and testing of hypotheses regarding its meaning. Rosci and colleagues<sup>22</sup> validated the PIT for detecting executive deficits in an Italian sample of 196 normal adults and 12 patients with pre-frontal brain lesions, who were asked to interpret what was happening in a reproduction of the famous painting “Il Sorcio” (“The Mouse”). Findings revealed that 60 percent of the patients were unable to interpret the picture. Moreover, a similar failure rate was found in patient performance on a verbal fluency task and the Trail Making Test. These results suggest the potential of the Italian version of the PIT for testing of pre-frontal patients, thus making it one of the most used neuropsychological tests in the Italian context<sup>25</sup>.

The current study was aimed to evaluate a 360° version of the PIT for detecting executive deficits through a function-led approach that combines experimental control with real-world engaging background. The study included patients with Parkinson's Disease (PD) because of the substantial research findings revealing a cognitive profile characterized by a dysexecutive syndrome<sup>6,26,27</sup>.

To investigate the quality of the experience associated with the experiencing the PIT 360°, we examined participant self-reports (e.g., perceived balance between challenge and skills, as well as patients' intrinsic motivation in being confronted with the task). To specifically investigate the capability of PIT 360° for detecting executive deficits, we compared the performance of patients with PD to that of healthy controls (HC) using indices obtained from PIT 360°. Furthermore, we compared performance between the two groups on traditional construct-driven neuropsychological assessments of executive functioning. Finally, to evaluate the predictive validity of indices obtained from PIT 360°, we investigated how all these measures would be able to distinguish patients with PD from HC into their respective groups.

## Results

**User experience assessment.** Table 1 presents descriptive data obtained from the user experience assessment divided between the two groups.

Comparison of data between groups using the Mann–Whitney U test did not reveal statistically significant differences (all  $ps > 0.05$ ). Results obtained from the Friedman Test indicated a significant difference among the four quadrants of GEW in terms of the mean number of reported felt emotions [ $\chi^2(3) = 87.572$ ;  $p < 0.001$ ] and their intensities [ $\chi^2(3) = 91,377$ ;  $p < 0.001$ ]. Specifically, Wilcoxon tests on mean number of self-reported emotion within the different quadrants revealed that all participants experienced more emotions with positive valence and high goal conduciveness. The same findings resulted for the intensities of self-reported emotions. See Tables 2 and 3 for full statistics.

**Conventional neuropsychological assessment.** Table 4 offers an overview of the scores obtained from the traditional neuropsychological evaluation divided between HC and patients with PD.

When controlling for age and education, results showed that the PD group had significantly lower scores on MoCa when compared to HC [ $F(1, 34) = 10.252$ ;  $p = 0.003$ ; Partial  $\eta^2 = 0.232$ ], they also performed significantly poorer on the phonemic fluency task [ $F(1, 34) = 6.390$ ;  $p = 0.016$ ; Partial  $\eta^2 = 0.158$ ] and in the two TMT sub-tests, namely the TMT-A [ $F(1, 34) = 7.075$ ;  $p = 0.012$ ; Partial  $\eta^2 = 0.172$ ] and the TMT-B [ $F(1, 34) = 4.240$ ;  $p = 0.047$ ; Partial  $\eta^2 = 0.111$ ].

**Performance on PIT 360°.** Fourteen patients with PD (73.7%) correctly interpreted the scene proposed in the PIT 360°, while 5 patients with PD (26.3%) failed in the recognition. As concerns the HC group, sixteen participants (84.2%) recognized the scene, while 3 participants (15.8%) didn't succeed in the task. There was no significant difference in the proportion of participants successfully completing the task between groups [ $\chi^2(1) = 0.426$ ;  $p = 0.693$ ].

When controlling for age and education, results showed that patients took longer (mean = 106.418; SD = 66.851) in comparison with HC group (mean = 70.808; SD = 52.782) for giving an interpretation of

	PD Group	HC Group
GEW 1 - Positive Valence/High coping potential Number of emotion	2.368 (1.165)	2.842 (1.167)
GEW 1 - Positive Valence/High coping potential Intensity	3.303 (0.975)	3.575 (1.186)
GEW 2 - Positive Valence/Low coping potential Number of emotion	0.737 (0.872)	1.316 (1.455)
GEW 2 - Positive Valence/Low coping potential Intensity	1.210 (1.484)	1.570 (1.465)
GEW 3 - Negative Valence/Low coping potential Number of emotion	0.105 (0.315)	0.053 (0.230)
GEW 3 - Negative Valence/Low coping potential Intensity	0.210 (0.713)	0.053(0.230)
GEW 4 - Negative Valence/High coping potential Number of emotion	0.0	0.053 (0.230)
GEW 4 - Negative Valence/High coping potential Intensity	0.0	0.158 (0.688)
Perceived coping skills (Flow Short Scale)	3.368 (0.831)	3.158 (0.688)
Perceived challenge (Flow Short Scale)	2.684 (0.671)	2.474 (0.697)
Perceived challenge- skill balance (Flow Short Scale)	2.474 (0.612)	2.421 (0.692)
IMI	2.947 (0.553)	3.189 (0.932)
SUS	4.1754 (1.496)	4.491 (1.517)

**Table 1.** Scores obtained from the user experience assessment for Parkinson's Disease patient (PD Group) and Older Controls (HC Group). Data are shown as means and standard deviations (SD).

			U	p
GEW 1	2.065 (1.175)	vs. GEW 2	-4.944	<0.001
		vs. GEW 3	-5.218	<0.001
		vs. GEW 4	-5290	<0.001
GEW 2	1.026 (1.219)	vs. GEW 3	-3.082	<0.001
		vs. GEW 4	-3.872	<0.001
GEW 3	0.079 (0.273)	vs. GEW 4	-1.000	0.317
GEW 4	0.026 (1.622)			

**Table 2.** Results obtained from Wilcoxon Test comparisons on different quadrants of Geneva Emotion Wheel. Mean number of reported felt emotion.

the scene proposed [ $F(1,34) = 4.624$ ;  $p = 0.039$ ; Partial  $\eta^2 = 0.120$ ]. Moreover, the PD group (mean = 10.00, SD = 7.039) provided a more detailed description of the scene in comparison to the HC group (mean = 5.789; SD = 3.505) [ $F(1,34) = 5.695$ ;  $p = 0.023$ ; Partial  $\eta^2 = 0.143$ ].

**Correlations between neuropsychological tests and performance on PIT 360°.** There were no significant correlations between neuropsychological tests and indexes of PIT 360° for HC group (see Fig. 1). For patients with PD, the time needed for a correct interpretation of the PIT 360° (Correct Interpretation) was found to positively correlate with performance on TMT-A ( $r = 0.509$ ;  $p = 0.026$ ) and negatively with that on the phonemic fluency task ( $r = -0.577$ ;  $p = 0.009$ ) (see Fig. 1). At the same time, there was a trend for correlation between Correct Interpretation and TMT-B ( $r = 0.429$ ;  $p = 0.066$ ).

**Classification of Healthy Controls or Patients with PD.** Performance of the classifiers was evaluated by carrying out a relative operating characteristic (ROC) analysis<sup>28</sup>. The area under the ROC curve (AUC) provides a single measure of overall prediction accuracy, *Precision* represents the proportion of true positives among all the instances classified as positive, *CA* (Classification accuracy) represents the proportion of the instances that were classified correctly, *F1* indicates the harmonic mean of precision (P) and Recall (R), and *Recall* (R) is the proportion of cases which were classified as positive, among all instances which truly were positives. Results from nonlinear stochastic approximation (i.e., machine learning approach) methods showed a Precision between 55.6% and 68.8% for the conventional neuropsychological assessment of executive functions (Table 5), while it ranged from 50.0% to 71.4% for PIT 360° (Table 6). According to these results (Tables 5 and 6), Random Forest has a Precision under the 60%, thus making this algorithm not reliable for the classification of cases into two groups using both traditional neuropsychological tests and indices from PIT 360°. On the other hand, results obtained with the Logistic Regression showed a good Precision only for traditional neuropsychological tests, but not for PIT 360°. Finally, we opted to use both Support Vector Machine and Naïve Bayes that showed a good precision (over than 60%).

Although the ability to predict control group membership is quite similar between the two types of assessment (slightly better for the traditional neuropsychological assessment), it is interesting to note that Naïve Bayes and Support Vector Machine algorithms showed that the indices from PIT 360° had a higher capability in predicting PD Group membership (see Fig. 2).

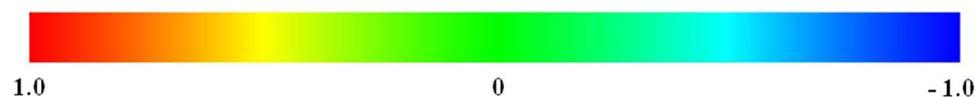
			U	p
GEW 1	3.439 (1.080)	vs. GEW 2	-5.105	<0.001
		vs. GEW 3	-5.176	<0.001
		vs. GEW 4	-5.249	<0.001
GEW 2	1.390 (1.466)	vs. GEW 3	-3.561	<0.001
		vs. GEW 4	-3.939	<0.001
GEW 3	0.131 (0.528)	vs. GEW 4	-5.577	0.577
GEW 4	0.080 (0.487)			

**Table 3.** Results obtained from Wilcoxon Test comparisons on different quadrants of Geneva Emotion Wheel. Intensities of reported felt emotion.

	PD Group	HC
Montreal Cognitive Assessment (MoCa)	26.000 (1.732)	27.789 (2.149)
Trail Making Test (TMT-A)	51.316 (24.637)	35.684 (17.327)
Trail Making Test (TMT-B)	135.824 (108.449)	80.263 (43.252)
F.A.S. Verbal Fluency	34.158 (9.800)	40.684 (9.939)

**Table 4.** Scores obtained from the neuropsychological assessment for Parkinson’s Disease patient (PD Group) and Older Controls (HC). Data are shown as means and standard deviations (SD).

PIT (seconds)		0.787 [<0.001]	0.185 [n.s.]	-0.095 [n.s.]	0.211 [n.s.]	-0.413 [n.s.]
PIT (NE)	0.768 [<0.001]		0.006 [n.s.]	-0.141 [n.s.]	0.296 [n.s.]	-0.117 [n.s.]
TMT-A	0.509 [0.026]	0.595 [n.s.]		0.857 [<0.001]	0.172 [n.s.]	-0.346 [n.s.]
TMT-B	0.429 [0.066]	0.128 [n.s.]	0.852 [<0.001]		-0.175 [n.s.]	-0.130 [n.s.]
FAS	-0.577 [0.009]	-0.213 [n.s.]	-0.593 [0.007]	-0.564 [0.012]		-0.042 [n.s.]
MoCA	-0.321 [n.s.]	-0.100 [n.s.]	-0.397 [n.s.]	-0.496 [0.030]	0.223 [n.s.]	
HC / PD	PIT (seconds)	PIT (NE)	TMT-A	TMT-B	FAS	MoCA



**Figure 1.** Correlation between neuropsychological tests and performance on PIT 360°. Correlation coefficients are represented by a color, which ranges from red (1) to blue (-1). Correlations between neuropsychological tests and indexes of PIT 360° for healthy controls (HC) are shown in the upper right, while findings for patients with Parkinson’s Disease (PD) are displayed in the bottom left.

### Discussion

The main objective of the current study was to evaluate the 360° version of the Picture Interpretation Test (PIT)<sup>22,23</sup>, for providing assessment of executive functions processing in PD using an innovative and ecologically valid tool. Following immersion in the PIT 360°, HCs and patients with Parkinson’s Disease (PD) were surveyed on their affective reactions to the 360° scene, perceived level of challenge and skills, appreciation of the tool, and their sense of presence while immersed in the PIT 360°. Then, to specifically evaluate the ability of

Method	AUC	CA	F1	Precision	Recall
Logistic Regression	0.658	0.658	0.629	0.668	0.579
Random Forest	0.533	0.533	0.541	0.556	0.526
Support Vector Machine (SVM)	0.658	0.658	0.667	0.650	0.684
Naïve Bayes	0.632	0.632	0.632	0.632	0.632

**Table 5.** Stratified 10-fold Cross validation for the neuropsychological assessment battery<sup>1</sup>. <sup>1</sup>AUC (Area under the ROC curve) is the area under the classic receiver-operating curve; CA (Classification accuracy) represents the proportion of the examples that were classified correctly; F1 represents the weighted harmonic average of the precision and recall (defined below); Precision represents a proportion of true positives among all the instances classified as positive. In our case, the proportion of condition correctly identified; Recall represents the proportion of true positives among the positive instances in our data.

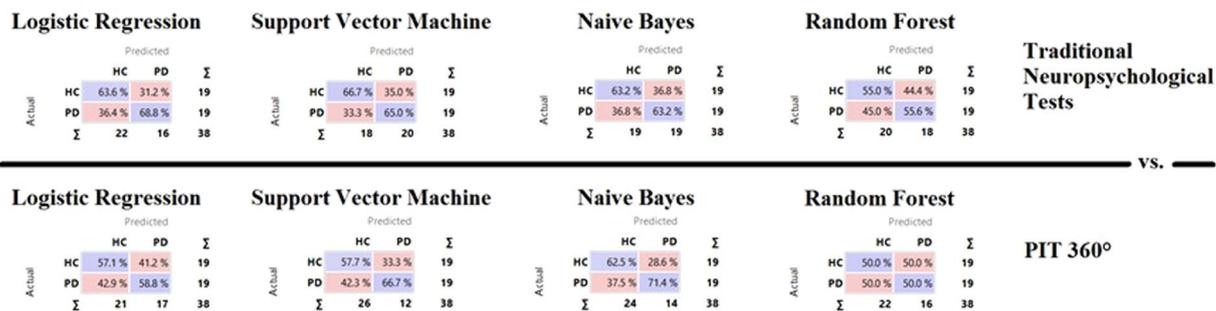
Method	AUC	CA	F1	Precision	Recall
Logistic Regression	0.579	0.579	0.556	0.588	0.526
Random Forest	0.500	0.500	0.457	0.500	0.421
Support Vector Machine (SVM)	0.605	0.605	0.516	0.617	0.421
Naïve Bayes	0.658	0.658	0.606	0.714	0.526

**Table 6.** Stratified 10-fold Cross validation for the indices of PIT 360°<sup>1</sup>. <sup>1</sup>AUC (Area under the ROC curve) is the area under the classic receiver-operating curve; CA (Classification accuracy) represents the proportion of the examples that were classified correctly; F1 represents the weighted harmonic average of the precision and recall (defined below); Precision represents a proportion of true positives among all the instances classified as positive. In our case, the proportion of condition correctly identified; Recall represents the proportion of true positives among the positive instances in our data.

PIT 360° in detecting executive deficits, we compared the performance of patients with PD and healthy controls comparing conventional neuropsychological assessments with the PIT 360°. Correlations between the conventional neuropsychological tests of executive functions and performance on PIT 360° were also explored. Finally, we investigated the predictive validity of indices obtained from PIT 360° in distinguishing PD patients from the healthy controls. Results from user experience assessment of the PIT 360° revealed that all participants endorsed positive reactions to their experience of the PIT 360°. This was apparent in the higher scores in the first quadrants of Geneva Emotion Wheel (GEW<sup>29,30</sup>), which includes interest, joy, happiness, satisfaction, elation and pride. In particular, patients with PD did not endorse affective responding with low valence and high control (such as anger or irritation). Moreover, both groups perceived a high level of their own skills in the context of a demanding task (the interaction with PIT 360°), which resulted in perceived balanced level of challenge-skills. As emerged by mean scores of Intrinsic Motivation Inventory (IMI<sup>31</sup>), the PIT 360° was reported to be an interesting and an enjoyable activity. Finally, all participants reported a very high level of presence during the interaction with PIT 360°.

All subjects were preliminarily assessed by a neuropsychological assessment and all of them obtained scores within the normal range. This confirms that our patients were in a relatively well-preserved clinical state. Only a statistical group comparison revealed differences between the two groups. As expected, these behavioral findings indicate that patients with early PD and no clinical evidence of cognitive impairment may already exhibit sub-clinical abnormalities, as previously reported<sup>32,33</sup>. In line with the pattern of results from conventional neuropsychological assessment, the PIT 360° analysis revealed different performances in patients with PD compared to HC. Although the percentage of PD patients that failed in correctly interpreting the scene is quite similar to that of HC (26.3% vs. 15.8%) confirming that they showed a relatively well-functioning cognitive status, analyses on the two PIT 360° indices showed significant differences between the two groups. Specifically, patients with PD took longer to provide a correct interpretation of the scene proposed and provided significantly more details about the objects found in the scene. While the patients gave significantly richer descriptions of the scene, they appeared more prone to distractor interference (“There is a white coat, there are two chairs in front of the TV. There is a big TV. On the floor, there is maybe a scale. Then, I see a wardrobe that may be a fridge. There is a man who is working on jumper cables near the wardrobe. I see an electric device on the table. I think that the man is repairing something. The man is curled up behind a white-board, or something similar, a spot where it is possible to hang sheets.”). These findings are in line with Luria’s view<sup>24</sup>, suggesting that this test is able to capture deficits in active visual perception. Our data indicated that patients with PD demonstrated more difficulties when compared to healthy controls in focusing on the most important components for a correct interpretation of the scene. Thus, PD patients appear to focus on details instead on the most informative elements. They were not able to find important elements for a correct interpretation of the whole scene nor did they match these elements with their hypotheses about the meaning. In most cases, a poor interpretation based only on the details was given.

Interestingly, results from correlation analyses indicated that neuropsychological tests correlate significantly with indexes of PIT 360° only for patients with PD. Specifically, the time needed for giving a correct interpretation of the PIT 360° scene did not correlate with the Montreal Cognitive Assessment (i.e., a measure of global cognitive level), but it was significantly correlated with the Trail Making Test and the phonemic verbal fluency



**Figure 2.** Classification of Healthy Controls or Patients with PD. The diagonal values (i.e., purple values) represent the elements for which the predicted group is equal to the true group, while off-diagonal elements are those that are mislabeled by the classifier. Naïve Bayes and Support Vector Machine algorithms demonstrated that PIT 360° has a higher capability in predicting PD Group membership with respect to traditional neuropsychological tests of executive functioning

task, thus tapping both verbal and visuospatial aspects of executive functioning and motor aspects. These findings suggest that PIT 360° can be considered as a quick, ecological and useful screening instrument able to evaluate different aspects of dysexecutive impairment in patients with PD.

Results obtained from classifiers clearly indicated the potential of PIT 360° scene assessment in distinguishing between patients with PD and HC. Two of the algorithms used indicated that PIT 360° had a higher capability in predicting PD group membership with respect to a traditional neuropsychological assessment. Although machine learning approaches have been traditionally applied to the analysis of very complex medical datasets<sup>34</sup>, recent studies have also applied them for classifying patients according to their cognitive impairment and consequently reduce the number of onerous tests required for their diagnosis<sup>35–37</sup>.

While the findings of the current study are promising, there are some limitations that should be considered. First of all, in order to fully evaluate the potentiality of PIT 360° as a new screening tool of executive functions, future studies are needed to assess its test–retest reliability and validity. A large validation study with a sample of participants across the lifespan including the PIT 360°, the original PIT<sup>22,23</sup>, as well as other conventional neuropsychological measures should be performed. Moreover, it will be important to investigate the value of PIT 360° in detecting executive impairments in other clinical populations who are known to have executive dysfunctioning.

## Conclusions

This study provides the first evidence that the 360° technology may play a role in the future of neuropsychological assessment. Moreover, this technology may be integrated with other portable devices, such as an eye-tracker. As suggested by pioneering study of Luria<sup>24</sup>, it would be particularly interesting to investigate patients' eye movements during the interpretation of the scene proposed. Indeed, Luria found that disturbances in the active visual perception were reflected by a corresponding disorganized scanning gaze movements. In conclusion, although preliminary, our findings provide encouraging evidence in support of the use of immersive 360° environments in general, and the PIT 360° specifically for innovative evaluation of executive impairments.

## Materials and Methods

**Participants.** Thirty-eight participants took part in the study: 19 patients with Parkinson's Disease (PD group), and 19 healthy controls matched for age and education with the PD group (HC group).

Outpatients meeting the diagnostic criteria for probable PD<sup>38</sup> were consecutively recruited from the Neurorehabilitation Unit of Don Carlo Gnocchi Foundation, IRCCS. All patients were at a mild to moderate stage of the disease, scoring between stages 1 and 2 of the Hoehn and Yahr (H&Y) Scale<sup>39</sup>. None had any report of cognitive problems or any evidence of cognitive deficits in their daily living activities. A Mini Mental State Examination (MMSE) was used to exclude any patient who reported scores outside the normal range (MMSE cut-off score 23, 8<sup>40,41</sup>). All subjects (patients and HCs) were right handed as assessed by the Edinburgh Inventory<sup>42</sup>. Exclusion criteria included any major systemic, psychiatric, or other neurological illnesses. Particular attention was used to exclude those patients who experienced visual hallucinations, had episodes of severe depression or autonomic failure, manifested resistance to dopaminergic drugs and were at an unstable dosage of antiparkinsonian treatment during the 3 months prior to study entry. The PD group was composed of 3 women and 16 men, while the HC group included 9 women and 10 men. The mean age for the PD group was 66.53 (SD = 9.43), with an average of 12.47 (SD = 3.47) years of education of, while the mean age for the HC group was 67.58 (SD = 7.86), with an average of 14.37 (SD = 3.48) years of education. The two groups did not differ significantly in terms of age [ $t(36) = -0.377$ ;  $p = 0.708$ ] or education [ $t(36) = -1.680$ ;  $p = 0.102$ ]. However, there were significantly less women in the PD group [ $\chi^2(1) = 4.835$ ;  $p = 0.036$ ].

The study was conducted in compliance with the Helsinki Declaration of 1975, as revised in 2008. Local Ethics Committee (Don Carlo Gnocchi Foundation) approval and written informed consent to be included in the study was obtained by participants before study initiation.



**Figure 3.** PIT 360°. Familiarization phase.

**Procedure of the study.** Participants underwent a conventional neuropsychological assessment to obtain their global cognitive profile and level of executive functioning (pre-task evaluation). Subsequently participants were asked to complete the PIT 360° (PIT 360° session). The PIT 360° was designed and administered through an innovative mobile application (PIT 360) that allows participants to explore an immersive 360° experience. At the end of PIT 360 task evaluation, subjects were asked to rate their affective reactions, perceived levels of challenge and skill, appreciation and sense of presence experienced while performing the PIT 360 task (post-task evaluation).

*Pre-task evaluation: neuropsychological measures.* In the pre-task evaluation we administered the following conventional paper and pencil tests: the Montreal Cognitive Assessment (MoCa<sup>43</sup>) as a measure of global cognitive level; the Trail Making Test (in two specific sub-tests: TMT-A and TMT-B<sup>44</sup>) and the phonemic verbal fluency task (FAS<sup>45</sup>) as measures of executive functioning.

*PIT 360° session: The PIT 360° development, description and administration.* The PIT 360° is the 360° version of the Picture Interpretation Test<sup>22,23</sup>. In the PIT test, a small-scale color reproduction (19 × 13) of the famous painting “Il Sorcio” (“The Mouse”) of the Italian painter Giacomo Favretto is used as a test stimulus. In this painting, a room in disarray is presented with three frightened girls standing on chairs and a boy who is searching for something on the floor. Although not visible, it is apparent that there is mouse hidden behind a piece of furniture. Participants are asked to interpret what is happening in the scene in a limited time frame (180 seconds), while the time to say the word “mouse” is the outcome measure.

The PIT 360° was developed with the Ricoh Theta S Digital Camera that permits the creation of 360° spherical imageries. The camera is able to capture a 360° scene by stitching two 180° scans via integrated software at a resolution of 1792 by 3584 pixels. This allows for a presentation of an immersive stereoscopic 360° experience directly on a virtual reality headset (including mobile phone) via the Ricoh Theta S application. For the present study, the PIT 360° was rendered through the mobile application of the Ricoh Theta S on an iPhone 6 Plus. Two scenes were recorded: one, to be used in the Familiarization phase (see Fig. 3), in which a meeting room appeared with tables, chairs, a sink, a television, some dressers with several objects spread on them. The second scene (see Fig. 4), to be used in experimental phase, was designed in line with the Favretto’s painting “Il Sorcio”: in the same room, with the same furniture and objects spread throughout it along with a boy searching for something on the floor, while three frightened girls standing on chairs watch him.

The neuropsychologist started the administrations by inviting participants to sit on a swivel chair and to wear the virtual reality headset (connected to the iPhone 6 Plus). This allowed participants to explore an interactive 360° experience. In case of presbyopia, participants were asked to wear their own glasses. Then participants underwent a familiarization phase (3 minutes) aimed at familiarizing them with the technology and control for potential side effects (e.g., dizziness, nausea). The examiner followed a cessation rule in which experimental sessions should be stopped if severe side effects occurred. The examiner asked participants to keep their eyes closed, and started time registration (in seconds) and audio recording coinciding with the instruction “Open your eyes”. Participants were then presented with the 360° scene of the room including a table (in the center), a sink with a mirror (on the participant’s right), a television on a table, two dressers (on the participant’s left), and various chairs and objects spread throughout the room. They were asked to find five objects in the scene to answer the experimenters’ questions (i.e. “Let’s search for the agenda. Where is the agenda?”). Upon completion of the three-minute familiarization phase, participants were asked to close again their eyes. The experimental session began with time registration (in seconds) and audio recording coinciding with the examiner’s instruction “Open your eyes”. In this phase, participants were asked to freely explore the scene derived from Favretto’s “Il Sorcio” and to tell the examiner what is happening as quickly as possible (maximum time: 180 seconds). Time registration lasted until the instant in which the participant said the word “mouse” or something similar (e.g., “snake”, “roach”, etc; generic classifications were allowed). After participants pronounced the word “mouse” (or similar generic classifications), the experimenter asked “What do you mean?” in order to confirm the participant’s understanding of the situation.



**Figure 4.** PIT 360°. Testing phase.

The following indices were calculated:

- 1) *Correct Interpretation*: The time in seconds registered from the time in which the experimenter said the words “Open your eyes” until the participant provided a correct interpretation” (i.e., “mouse”, “animal”, etc.). The maximum time allowed was 180 seconds. If the participants failed to interpret the scene in the allotted 180 seconds, then a time of 180 seconds was assigned as the outcome (as suggested by Rosci and colleagues<sup>22</sup>);
- 2) *Number of Scene Elements*: The sum of the scene elements that were verbalized during the interpretation of the scene.

*The post-task evaluation: user experience assessment.* In the post-task evaluation, participants were asked to rate their experience during the task on the following instruments:

Geneva Emotion Wheel (GEW<sup>29,30</sup>).

It consists of 20 discrete emotion terms that are systematically aligned in a circle. Underlying the alignment of the emotion terms are the two dimensions – valence/goal conduciveness (negative to positive) and control/coping potential (low to high) separating the emotions into four quadrants, each meant as an emotion family: Positive valence/High coping potential, Positive valence/Low coping potential, Negative valence/Low coping potential, and Negative valence/High coping potential. In all the four quadrants, the single emotion terms are considered as indexes “reflecting a unique experience of mental and bodily changes in the context of being confronted with a particular event”<sup>29</sup>. The mean number of emotions labels chosen within each of the four quadrants and the reported intensity in feeling it show how participants shaped their subjective feeling in performing the PIT 360° task along the dimensions of valence/goal conduciveness and control/coping potential.

Perceived fit of demands and skills (from Flow Short Scale<sup>46</sup>). Landhäuser and Keller<sup>47</sup> highlighted that, although researchers seem to equalize the skill-demands compatibility with the experience of flow itself (e.g. <sup>48,49</sup>) in many studies, the balance between skills of the individual and perceived challenges of the task cannot be considered as a measure of the flow experience per se. Therefore, we administered the three items from the Flow Short Scale (5-points scale) that assess this specific component of the flow experience in performing the PIT 360° task. The first item asked participants to evaluate their perceived level of skills in coping with the task (“Perceived coping skills”), whereas the second item is related to the perceived level of challenges (“Perceived challenge”). Finally, participants were asked to indicate the perceived challenge-skill balance (“Perceived challenge- skill balance”) in 5-point-scale with 1 indicating that the current challenge is too low for ones’ perceived skills, 3 indicating that the current challenge fit exactly to ones’ skills and 5 indicating that the challenge is too high.

Intrinsic Motivation Inventory (IMI<sup>31</sup>). Participants responded to five items (7-points scale) from the sub-scale “Enjoyment” of the Intrinsic Motivation Inventory (IMI). These items were chosen to evaluate participants’ appreciation to the proposed activity, including the items “This activity was fun to do” and “While I was doing this activity, I was thinking about how much I enjoyed it”. The mean of the item scores is considered.

Slater-Usch-Steed Questionnaire (SUS<sup>50</sup>). It consists of 7-points questionnaire which evaluates the sense of presence with three items: 1) the sense of being in the scene depicted in the 360° scene, 2) the extent to which the 360° scene became the dominant reality, and 3) the extent to which the 360° scene was remembered as a place. The mean of the item scores was considered.

**Data analysis.** The normality of data distribution was assessed using the Kolmogorov-Smirnov test. Since data were not normally distributed, non-parametric tests were used to investigate the quality of the experience associated with the interaction with PIT 360° and potential differences between PD and HC group on user experience variables (i.e., GEW, Flow Short Scale, IMI, and SUS). Moreover, for the GEW<sup>29,30</sup>, differences within the four quadrants in the number and intensity of self-reported emotions were also investigated using the Friedman

Test. Next, Wilcoxon tests, with Bonferroni's adjustment, were computed to break down significant findings. Subsequently, between-group comparisons of performance on the conventional neuropsychological assessment and the indexes of PIT 360° (Correct Interpretation and Number of Scene Element) were made by univariate analysis of covariance (ANCOVA), using age and education as covariates. These statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (IBM Corp Armonk, NY, USA), version 23. Finally, Pearson correlation coefficient was used to examine correlations between conventional neuropsychological tests and the indexes of PIT 360°. These statistical analyses were carried out using the software MedCalc (MedCalc Software, Ostend, Belgium), version 16.8.4.

Nonlinear stochastic approximation (i.e., machine learning) methods were used to compare the classification accuracy of traditional neuropsychological assessments versus the PIT 360° indices for classifying participants into either the “Patients with PD” or “Healthy Controls” groups. Machine learning approaches are devoted to prediction and it is thought to be explorative rather than explicative<sup>51</sup>; accordingly, we used different algorithms to compare the predictive value of each one of them to understand which one was the best based on their accuracy. Because our analyses are based on relatively small sample sizes, a Leave-one-out cross-validation (LOOCV)<sup>52,53</sup>. Different algorithms were employed, namely: a) a Logistic Regression classification algorithm with ridge regularization; b) a Random Forest classification to classify using an ensemble of decision trees; c) a Support Vector Machine (SVM) to map inputs to higher-dimensional feature spaces that best separate different classes; d) a Naïve Bayes classification, for discriminating between the two groups, even without any particular assumption for the distribution for the features. All these analyses were computed using Python 3.4 with the Orange 3.3.5 data mining suite, which is freely available as open source code (<https://github.com/biolab/orange3>).

## References

- Stuss, D. T. & Alexander, M. P. Executive functions and the frontal lobes: a conceptual view. *Psychological research* **63**, 289–298 (2000).
- Barker, L. A., Andrade, J. & Romanowski, C. A. J. Impaired implicit cognition with intact executive function after extensive bilateral prefrontal pathology: A case study. *Neurocase* **10**, 233–248 (2004).
- Goldstein, G. In *Ecological Validity of Neuropsychological Testing* (eds R. J. Sbordone & C. J. Long) 75–89 (FL:GSRPress/St.LuciePress, 1996).
- Shallice, T. & Burgess, P. W. Deficits in strategy application following frontal lobe damage in man. *Brain* **114**, 727–741 (1991).
- Chaytor, N. & Schmitter-Edgecombe, M. The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology review* **13**, 181–197 (2003).
- Kudlicka, A., Clare, L. & Hindle, J. V. Executive functions in Parkinson's disease: Systematic review and meta-analysis. *Movement Disorders* **26**, 2305–2315 (2011).
- Spooner, D. M. & Pachana, N. A. Ecological validity in neuropsychological assessment: a case for greater consideration in research with neurologically intact populations. *Archives of clinical neuropsychology* **21**, 327–337 (2006).
- Burgess, P. W. *et al.* The case for the development and use of “ecologically valid” measures of executive function in experimental and clinical neuropsychology. *Journal of the international neuropsychological society* **12**, 194–209 (2006).
- Parsons, T. D. Virtual Reality for Enhanced Ecological Validity and Experimental Control in the Clinical, Affective, and Social Neurosciences. *Frontiers in Human Neuroscience*, 1–19 (2015).
- Chan, R. C. K., Shum, D., Touloupoulou, T. & Chen, E. Y. H. Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of clinical neuropsychology* **23**, 201–216 (2008).
- Lezak, M. D. *Neuropsychological Assessment (4th Ed.)*. (Oxford University Press, 2004).
- Lehto, J. E., Juujärvi, P., Kooistra, L. & Pulkkinen, L. Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology* **21**, 59–80 (2003).
- Miyake, A. *et al.* The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive psychology* **41**, 49–100 (2000).
- Diamond, A. Executive functions. *Annual review of psychology* **64**, 135 (2013).
- Logie, R. H., Trawley, S. & Law, A. Multitasking: Multiple, domain-specific cognitive functions in a virtual environment. *Memory & cognition* **39**, 1561–1574 (2011).
- Parsons, T. D., Carlew, A. R., Magtoto, J. & Stonecipher, K. The potential of function-led virtual environments for ecologically valid measures of executive function in experimental and clinical neuropsychology. *Neuropsychological rehabilitation*, **37**(5), 777–807 (2017).
- Bohil, C. J., Alicea, B. & Biocca, F. A. Virtual reality in neuroscience research and therapy. *Nature reviews neuroscience* **12**, 752–762 (2011).
- Raspelli, S. *et al.* Validating the Neuro VR-based virtual version of the Multiple ErrandsTest: preliminary results. *Presence: Teleoperators and Virtual Environments* **21**, 31–42 (2012).
- Cipresso, P. *et al.* Virtual multiple errands test (VMET): a virtual reality-based tool to detect early executive functions deficit in Parkinson's disease. *Frontiers in behavioral neuroscience* **8** (2014).
- Cipresso, P. *et al.* Break in volition: A virtual reality study in patients with obsessive-compulsive disorder. *Experimental brain research* **229**, 443–449 (2013).
- Milgram, P. & Kishino, F. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems* **77**, 1321–1329 (1994).
- Rosci, C., Sacco, D., Laiacona, M. & Capitani, E. Interpretation of a complex picture and its sensitivity to frontal damage: a reappraisal. *Neurological Sciences* **25**, 322–330 (2005).
- Bisiach, E., Cappa, S. & Vallar, G. *Guida all'esame neuropsicologico*. (R. Cortina, 1983).
- Luria, A. R., Karpov, B. A. & Yarbuss, A. L. Disturbances of active visual perception with lesions of the frontal lobes. *Cortex* **2**, 202–212 (1966).
- Bianchi, A. & Dai Prà, M. Twenty years after Spinnler and Tognoni: new instruments in the Italian neuropsychologist's toolbox. *Neurological sciences* **29**, 209–217 (2008).
- Dirnberger, G. & Jahanshahi, M. Executive dysfunction in Parkinson's disease: a review. *Journal of neuropsychology* **7**, 193–224 (2013).
- Litvan, I. *et al.* Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. *Movement Disorders* **27**, 349–356 (2012).
- Swets, J. A. Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293 (1988).
- Scherer, K. R. What are emotions? And how can they be measured? *Social science information* **44**, 695–729 (2005).
- Scherer, K. R., Shuman, V., Fontaine, J. R. J. & Soriano, C. The GRID meets the Wheel: Assessing emotional feeling via self-report. *Components of emotional meaning: A sourcebook*, 281–298 (2013).
- Deci, E. L., Eghrari, H., Patrick, B. C. & Leone, D. R. Facilitating internalization: The self-determination theory perspective. *Journal of personality* **62**, 119–142 (1994).

32. Pillon, B., Czernecki, V. & Dubois, B. Dopamine and cognitive function. *Current opinion in neurology* **16**, S17–S22 (2003).
33. Baglio, F. *et al.* Functional brain changes in early Parkinson's disease during motor response and motor inhibition. *Neurobiology of aging* **32**, 115–124 (2011).
34. Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine* **23**, 89–109 (2001).
35. Buscema, M. *et al.* Artificial neural networks and artificial organisms can predict Alzheimer pathology in individual patients only on the basis of cognitive and functional status. *Neuroinformatics* **2**, 399–415 (2004).
36. Parsons, T. D., Rizzo, A. A. & Buckwalter, J. G. Backpropagation and regression: comparative utility for neuropsychologists. *Journal of Clinical and Experimental Neuropsychology* **26**, 95–104 (2004).
37. Weakley, A., Williams, J. A., Schmitter-Edgecombe, M. & Cook, D. J. Neuropsychological test selection for cognitive impairment classification: A machine learning approach. *Journal of clinical and experimental neuropsychology* **37**, 899–916 (2015).
38. Gelb, D. J., Oliver, E. & Gilman, S. Diagnostic criteria for Parkinson disease. *Archives of neurology* **56**, 33–39 (1999).
39. Fahn, S. & Elton, R. L. Unified rating scale for Parkinson's disease. *Recent developments in Parkinson's disease*. Florham Park. New York: Macmillan, 153–163 (1987).
40. Measso, G. *et al.* The mini-mental state examination: Normative study of an Italian random sample. *Developmental Neuropsychology* **9**, 77–85 (1993).
41. Folstein, M. F., Robins, L. N. & Helzer, J. E. The mini-mental state examination. *Archives of general psychiatry* **40**, 812–812 (1983).
42. Oldfield, R. C. Handedness in musicians. *British Journal of Psychology* **60**, 91–99 (1969).
43. Conti, S., Bonazzi, S., Laiacona, M., Masina, M. & Coralli, M. V. Montreal Cognitive Assessment (MoCA)-Italian version: regression based norms and equivalent scores. *Neurological Sciences* **36**, 209–214 (2015).
44. Giovagnoli, A. R. *et al.* Trail making test: normative values from 287 normal adult controls. *The Italian journal of neurological sciences* **17**, 305–309 (1996).
45. Carlesimo, G. A. *et al.* The mental deterioration battery: normative data, diagnostic reliability and qualitative analyses of cognitive impairment. *European neurology* **36**, 378–384 (1996).
46. Rheinberg, F., Vollmeyer, R. & Engeser, S. In *Diagnostik von Motivation und Selbstkonzept* (eds J. Stiensmeier-Pelster & F. Rheinberg) 261–279 (2003).
47. Landhäuser, A. & Keller, J. In *Advances in flow research* (ed S. Engeser) 65–85 (Springer, 2012).
48. Csikszentmihalyi, M. & LeFevre, J. Optimal experience in work and leisure. *Journal of personality and social psychology* **56**, 815–822 (1989).
49. Eisenberger, R., Jones, J. R., Stinglhamber, F., Shanock, L. & Randall, A. T. Flow experiences at work: For high need achievers alone? *Journal of Organizational Behavior* **26**, 755–775 (2005).
50. Usoh, M., Catena, E., Arman, S. & Slater, M. Using presence questionnaires in reality. *Presence: Teleoperators and Virtual Environments* **9**, 497–503 (2000).
51. Mitchell, T. *Machine Learning.*, (McGraw Hill, 1997).
52. Caruana, R. & Niculescu-Mizil, A. In *Proceedings of the 23rd international conference on Machine learning*. 161–168 (ACM).
53. Suthaharan, S. In *Machine Learning Models and Algorithms for Big Data Classification* 183–206 (Springer, 2016).

## Acknowledgements

This work was partially supported by the Italian funded project “High-end and Low-End Virtual Reality Systems for the Rehabilitation of Frailty in the Elderly” (PE-2013-02355948), by the research project Tecnologia Positiva e Healthy Aging (Positive Technology and Healthy Aging) (Grant D.3.2., 2014) and by the research project “Ageing and Healthy Living: A Human Centered Approach in Research and innovation as Source of Quality Life”, funded by Fondazione Cariplo within the 2014. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication

## Author Contributions

S.S., F.B., F.R., O.R., and G.R. developed the study concept. All authors contributed to the study design. F.R. was involved in the data collection. G.C. was responsible for the technical development of the PIT. S.S. and O.R. performed the data analysis and interpretation under the supervision of F.M., R.M., G.D.L., T.P.C. performed computational data analysis. S.S., F.B., F.R., and O.R. wrote the first draft of the manuscript. All authors were involved in a critical revision of the manuscript for important intellectual content. All the authors approved the final version of the manuscript for submission.

## Additional Information

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017