RESEARCH ARTICLE

Unsupervised assessment of cognition in the Healthy Brain Project: Implications for web-based registries of individuals at risk for Alzheimer's disease

Stephanie Perin[1,2,3] | Rachel F. Buckley[1,4,5,6] | Matthew P. Pase[1,2,7] |
Nawaf Yassi[1,8,9] | Alexandra Lavale[1,2] | Peter H. Wilson[3] | Adrian Schembri[10] |
Paul Maruff[1,10] | Yen Ying Lim[1,2]

[1] Melbourne Dementia Research Centre, Florey Institute of Neuroscience and Mental Health and the University of Melbourne, Parkville, Victoria, Australia

[2] School of Psychological Sciences, Turner Institute for Brain and Mental Health, Monash University, Clayton, Victoria, Australia

[3] School of Psychology, Faculty of Health Sciences, Australian Catholic University, Melbourne, Victoria, Australia

[4] Melbourne School of Psychological Sciences, University of Melbourne, Parkville, Victoria, Australia

[5] Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

[6] Center for Alzheimer Research and Treatment, Department of Neurology, Brigham and Women's Hospital, Boston, Massachusetts, USA

[7] Harvard T.H. ChanSchool of Public Health, Boston, Massachusetts, USA

[8] Department of Medicine and Neurology, Melbourne Brain Centre at The Royal Melbourne Hospital, University of Melbourne, Parkville, Victoria, Australia

[9] Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia

[10] Cogstate Ltd, Melbourne, Victoria, Australia

Correspondence
Yen Ying Lim, School of Psychological Sciences, Turner Institute for Brain and Mental Health, Monash University, 18 Innovation Walk, Clayton, Victoria 3800, Australia.
E-mail: yenying.lim@monash.edu

Abstract

Introduction: Web-based platforms are used increasingly to assess cognitive function in unsupervised settings. The utility of cognitive data arising from unsupervised assessments remains unclear. We examined the acceptability, usability, and validity of unsupervised cognitive testing in middle-aged adults enrolled in the Healthy Brain Project.

Methods: A total of 1594 participants completed unsupervised assessments of the Cogstate Brief Battery. Acceptability was defined by the amount of missing data, and usability by examining error of test performance and the time taken to read task instructions and complete tests (learnability).

Results: Overall, we observed high acceptability (98% complete data) and high usability (95% met criteria for low error rates and high learnability). Test validity was confirmed by observation of expected inverse relationships between performance and increasing test difficulty and age.

Conclusion: Consideration of test design paired with acceptability and usability criteria can provide valid indices of cognition in the unsupervised settings used to develop registries of individuals at risk for Alzheimer's disease.

1 | INTRODUCTION

Over the past decade, studies of Alzheimer's disease (AD) pathogenesis have examined relationships between cognitive dysfunction and AD biomarkers, which has enabled the characterization of very early stages of the disease (eg, preclinical AD).[1–3] One key theme arising from these prospective clinicopathological studies is that very large samples are required to detect these subtle but important interrelationships in preclinical AD or in individuals at risk of developing AD.[4,5] Conventional approaches to identifying early AD-related cognitive dysfunction have required individuals to attend medical research facilities at regular (eg, annual) intervals and undergo detailed assessments.[6,7] However, such approaches are expensive, time consuming, and can be burdensome to participants.[8,9] These limitations have prompted development of cognitive tests relevant to preclinical AD which can be administered remotely and in an unsupervised manner, for example, using online platforms to develop registries of people who may be at increased risk of dementia.[10,11] Unsupervised cognitive testing allows for the rapid recruitment of large samples in a cost-effective manner, a reduction in administrator bias, automated scoring by computerized algorithms, and the ability to maximize sample generalizability by facilitating the inclusion of participants who reside in remote locations.[8,10–12]

While unsupervised cognitive testing facilitates the collection of large-scale data, the quality of this data may be impacted by variables that can otherwise be controlled in supervised settings such as external disruptions to testing.[13] The absence of an assessor to establish rapport may also cause the unsupervised assessment to become impersonal and thereby reduce levels of motivation or engagement leading to sample attrition.[9,10] Additionally, it cannot be presumed that cognitive tests applied in unsupervised settings are equivalent psychometrically to the same tests used in supervised settings.[8] In particular, without a supervisor it is difficult to determine whether poor performance reflects that the individual did not understand or comply with the rules or response requirements of the cognitive test, or whether it reflects a true reduction in the aspect of cognition measured by that test.[10] Thus, in unsupervised cognitive testing there is a need for indices of performance that may provide guidance on this decision.

Human Computer Interaction (HCI) is a field of research which explores the interaction between humans and computerized technology.[14] In particular, HCI research looks to study the user's experience (ie, usability and acceptability) of the computerized platform in question.[15] HCI models, therefore, provide a useful framework for examining the development of unsupervised tests administered remotely using technology.[16,17] The HCI concepts of acceptability and usability have been applied in experimental studies of supervised computerized cognitive tests.[18–20] In these studies, HCI acceptability was operationalized as the amount and nature of missing data, by determining whether particular tests were more likely to demonstrate missing data.[21] HCI usability was operationalized as the amount of time taken to read the test instructions and to complete each test (learnability) and participants' ability to adhere to the requirements of each test (error).[17,22] These same concepts can be applied to understand the quality of data from unsupervised cognitive tests administered through web-based platforms in large and unselected samples of individuals at risk of AD. It is then possible to investigate their validity using standard psychological approaches.

The first aim was to apply an HCI framework to determine the HCI acceptability and usability of unsupervised cognitive tests conducted on participants enrolled in the Healthy Brain Project (HBP), an online study of at-risk middle-aged adults with high rates of first-degree family history of dementia.[23] The first hypothesis was that unsupervised cognitive tests conducted via a web-based platform would have high HCI acceptability (low rates of missing data), and high HCI usability (high learnability; low error rates). The second aim was to determine the validity of unsupervised cognitive tests by examining the relationships between performance and test difficulty, and between performance and age.[24] The second hypothesis was that decreasing performance would be associated with increasing test difficulty, and that participants would show also expected age-related decreases in test performance. The third aim was to explore the effects of testing environment (eg, home/work alone, home/work with others around, public space), and first-degree family history of dementia on cognitive performance. The third hypothesis was that participants who complete testing in environments with higher levels of distraction (eg, with others present, or in a public space), and participants with a first-degree family history of dementia, will show worse cognitive performance than those who complete testing in a quiet environment (eg, at home, or alone), and those without a first-degree family history of dementia.

2 | METHODS

2.1 | Participants

The sample consisted of 1594 participants who had completed unsupervised cognitive testing as part of their enrollment in the HBP (healthybrainproject.org.au). Participants were aged 40 to 65 years, lived in the community, and self-reported family history of dementia. Participants were excluded from enrolment in HBP if they self-reported any of the following: history of major traumatic brain injury; diagnosis of AD, Parkinson's disease, Lewy body dementia, or any other type of dementia; previous use of medications for the treatment of AD; current use of narcotics or antipsychotic

medications; uncontrolled major depression, schizophrenia, or another Axis I psychiatric disorder described in the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition (DSM-IV) within the past year; or a history of alcohol or substance abuse or dependence within the past two years. Participants were recruited through a series of media appeals, community-based and scientific organizations (eg, Dementia Australia and the Florey Institute of Neuroscience and Mental Health), traditional word of mouth, and social media. Participants self-referred to the study. Further details regarding the inclusion and exclusion criteria and recruitment process for the HBP have been described elsewhere.[23] As recruitment into the HBP is ongoing, the current report only includes data that have been collected up to the first formal DataFreeze (August 2018).[23]

The HBP platform was designed to allow participants to complete the study questionnaires and surveys over multiple sessions, and in any order. However, participants were required to complete their cognitive testing in one sitting. Further, the cognitive tests are administered in a pre-specified order which the participant must follow. The aim of this approach was to reduce participant burden and to allow participants' maximum flexibility to participate in the study in their own time, while ensuring consistency in the administration of the cognitive tests. We have previously reported that participants typically complete all HBP assessments across a month.[23] The HBP was approved by the human research ethics committee of Melbourne Health.

RESEARCH IN CONTEXT

1. Systematic review: The authors reviewed the literature using traditional (eg, PubMed) sources, meeting abstracts, and presentations. Studies that reported on the remote or unsupervised assessment of cognitive tests using online or web-based platforms were included. Studies relating to the Human Computer Interface (HCI) framework were also reviewed.

2. Interpretation: Our findings suggest that HCI criteria for acceptability and usability provide a sound framework for the design and administration of cognitive tests in unsupervised settings. With this context operating, the psychometric characteristics of the cognitive data collected in unsupervised settings were equivalent to that collected from supervised assessment.

3. Future directions: We provide a framework for the evaluation of cognitive tests and surveys that are administered via online or web-based platforms. Further studies will be required to determine participant retention and attrition, as well as the test-retest reliability, and sensitivity of unsupervised cognitive assessments to detecting future dementia risk.

2.2 | Cognitive tests

Unsupervised cognitive testing was carried out using the Cogstate Brief Battery (CBB) for which instructions and delivery have been modified for online assessment.[19,25] The CBB has to be completed on a web browser (Internet Explorer, Google Chrome, or Safari), and as such, participants were required to complete their cognitive testing on a desktop or laptop. Participants were directed to the CBB platform via the HBP website. Upon launching the CBB platform, participants were required to complete the test battery in one sitting. The approximate completion time for the entire CBB is 20 minutes. CBB tests are downloaded onto the browser, completed locally, and then uploaded. This allowed for the minimization of the impact of internet connectivity on test performance. The CBB has a game-like interface which uses playing card stimuli and requires participants to provide "Yes" or "No" responses. The CBB consists of four tests: Detection (DET), Identification (IDN), One Card Learning (OCL), and One-Back (OBK). These tests have been described in detail previously,[26] and Figure 1 illustrates an example of the unsupervised assessment training procedure for the IDN test. Briefly, DET assesses psychomotor function, and IDN assesses visual attention. The primary outcome for both DET and IDN was reaction time in milliseconds (speed), which was normalized using a $\log_{10}$ transformation.[27] OCL assesses visual learning, and OBK assesses working memory and attention. The primary outcome measure for OCL and OBK was proportion of correct answers (accuracy), which was normalized using an arcsine square-root transformation.[27] The four different tests in the CBB were designed to use identical stimulus material, stimulus displays, and response requirements. With this consistency, the demands of the battery increase as each test in the battery is completed successfully. The CBB begins with the DET test, which requires a single button press in response to a stimulus change. The IDN test is presented second and this requires that in addition to detecting the stimulus change, individuals must also discriminate stimuli based on their color (red or black). The OBK is the third test and requires that discrimination decisions be based on color and number information held in working memory. The OCL is the final and most difficult test on the basis that it requires decisions be based on stimulus information learned and retained throughout the test itself. Consistent with Sternberg additive factors logic,[27] this increase in the complexity of cognitive operations required for optimal performance on the different CBB tests, manifests in performance as increasing response times and decreasing accuracy. Such relationships are observed reliably in supervised studies of cognitively normal younger[28] and older adults,[20] as well as in patients with AD in the preclinical, prodromal, and dementia stages.[27,29,30] As such, the examination of the same test difficulty/performance relationships would provide a sound criterion for determining test validity in unsupervised performance.

2.3 | Testing environment survey

Participants were asked to indicate (1) never, (2) rarely, (3) often, or (4) all the time to the following questions: type of environment typically
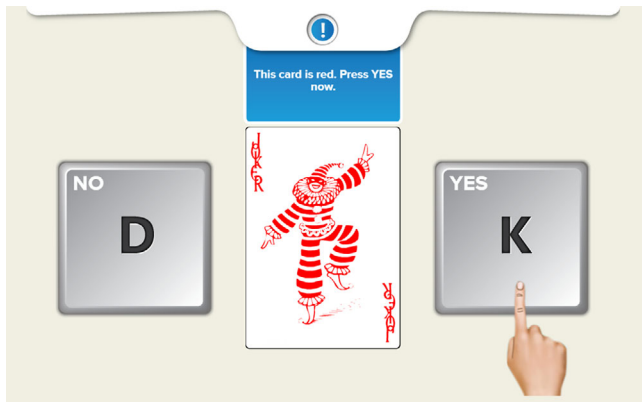
FIGURE 1 Example of training procedure interface for the Identification (IDN) test in the Cogstate Brief Battery (CBB)

in while doing the tests and surveys on the HBP, (1) at home without other people present, (2) at home with other people present (but are quiet), (3) at home with other people present (but are noisy), (4) at work without other people present, (5) at work with other people present (but are quiet), (6) at work with other people present (but are noisy), (7) during your commute, (8) in a public space (eg, public park, restaurant, café) that was quiet, and (9) in a public space (eg, public park, restaurant, café) that was noisy.

2.4 | Data analysis

All analyses were conducted using R Statistical Computing Software (R version 3.5.0). Given the large number of comparisons, statistical significance was adjusted using the false-discovery rate (FDR) correction.[31]

For each cognitive outcome, HCI acceptability was defined as participants' ability to complete tests in an unsupervised context and determined by examining the amount and nature of missing data, by determining whether particular tests in the CBB were more likely to demonstrate missing data. Missing data was defined as the non-completion of a test, or failure to provide responses to at least 75% of trials within a test. HCI usability was defined as (1) HCI error, operationalized by examining the nature and number of tests that do not meet the established error criterion, and (2) HCI learnability, operationalized by examining the mean and standard deviation (SD) of the time taken to read test instructions and complete each test, and also the number of participants who exceeded 2 SD from the average time taken to read test instructions and complete each test. The HCI error criterion for each test has been established previously[20,32] as an index of effort, motivation, and potential distraction, and is defined as: (1) 80% accuracy for the DET and IDN tests, (2) 50% accuracy for the OCL test, and (3) 70% accuracy for the OBK test. HCI error was considered to have occurred systematically for participants who met pre-specified error criteria on at least two cognitive tests. The validity of each cognitive test was determined by examining associations between (1) test performance and test difficulty, and (2) test performance and age. Participant age was treated both as a continuous and categorical

variable, where participants were grouped into (1) 40 to 49 years, (2) 50 to 59 years, and (3) 60 to 70 years.

Finally, in a subgroup of participants who also responded to a survey of their testing environment (n = 827), we explored whether cognitive performance varied across different testing environments. To achieve this, we grouped participants into (1) "alone" (ie, those who responded "all the time" or "often" to completing tests at home or at work without other people present), (2) "with others (quiet)" (ie, those who responded "all the time" or "often" to completing tests at home or at work with other people present but are quiet), and (3) "with others (noisy)"(ie, those who responded "all the time" or "often" to completing tests at home, work, public spaces, or during their commute where other people are present and are noisy).

3 | RESULTS

3.1 | Sample overview

The demographic characteristics of our sample are provided in Table 1. Older age groups were associated with lower levels of education, lower annual income, and fewer symptoms of depression and anxiety. Older age groups were also more likely to report a first-degree family history of dementia, and postcodes from rural/regional locations.

3.2 | HCI acceptability of unsupervised cognitive testing

Analyses of HCI acceptability indicated that 30 (1.9%) cases of missing data occurred for DET, and 10 (0.6%) cases of missing data occurred for IDN. No missing data occurred for the OCL or OBK.

3.3 | HCI usability of unsupervised cognitive testing

Analyses of HCI error indicated that the proportion of completed cognitive tests that did not exceed pre-specified error criteria were 93.4% for DET (n = 1488), 89.5% for IDN (n = 1427), 94.6% for OCL (n = 1508), and 94.7% for OBK (n = 1510). When this analysis was considered for the four cognitive tests simultaneously, 19.6% (n = 313) of participants exceeded error criteria for one cognitive test, 4.6% (n = 73) for two cognitive tests, 2.0% (n = 32) for three cognitive tests, and 1.6% (n = 25) on all four cognitive tests. Examination of the demographic characteristics of participants who did and did not meet HCI error criteria indicated that participants who did not meet error criteria on at least two cognitive tests were more likely to be male and required significantly longer time to complete all cognitive tests (Table 2).

Results from the analysis of HCI learnability (ie, time taken to read instructions and complete tests) are summarized in Table 3. As a significant difference in the amount of time taken to complete tests was observed between participants who passed or failed HCI error

TABLE 1 Sample demographic characteristics

	Overall	40 to 49 years	50 to 59 years	60 to 70 years	
N	1594	321	745	528	P
N (%) females	1184 (74.3%)	232 (72.3%)	575 (77.2%)	377 (71.4%)	.106
N (%) rural/regional location	399 (25.0%)	60 (18.7%)	173 (23.2%)	166 (31.4%)	<.001
N (%) white	1135 (71.2%)	231 (72.0%)	512 (68.7%)	392 (74.2%)	.106
N (%) first-degree dementia family history	737 (46.2%)	107 (33.3%)	361 (48.5%)	269 (50.9%)	<.001
Age	56.23 (6.60)	46.05 (2.82)	55.69 (2.76)	63.19 (2.07)	<.001
Education (y)	11.87 (3.40)	12.50 (3.24)	11.61 (3.40)	11.85 (3.47)	<.001
Annual income (self) ('000s)	65.56 (34.88)	76.22 (36.41)	67.46 (35.09)	56.20 (31.13)	<.001
HADS depression	2.56 (3.09)	3.05 (3.42)	2.58 (3.25)	2.23 (2.58)	.003
HADS anxiety	4.21 (3.30)	4.69 (3.57)	4.24 (3.36)	3.86 (2.98)	.004
No. of days between enrollment and test completion	42.39 (81.26)	47.93 (80.64)	44.10 (84.99)	36.63 (75.87)	.106

Note: All values for continuous variables (age, education, annual income, HADS depression, HADS anxiety, and no. of days between enrollment and test completion) presented as mean (standard deviation); chi-square was used to test differences between groups for categorical variables, and analysis of variance was used to test differences between groups for continuous variables; FDR corrections have been applied to all *P* values.
Abbreviations: FDR, false-discovery rate; HADS, Hospital Anxiety and Depression scale.

TABLE 2 Demographic characteristics of overall sample, and of participants who did and did not meet error criteria on at least two cognitive tests

	Overall	Failed error[a]	Passed error	
N	1594	73	1521	P
N (%) females	1184 (74.3%)	48 (65.8%)	1136 (74.7%)	<.001
N (%) rural/regional location	399 (25.0%)	16 (21.9%)	383 (25.2%)	.606
N (%) white	1135 (71.2%)	47 (64.4%)	1088 (71.5%)	.301
Age	56.23 (6.60)	57.74 (6.59)	56.16 (6.60)	.120
Education	11.87 (3.40)	11.03 (3.21)	11.91 (3.41)	.120
Annual income (self) ('000s)	65.56 (34.88)	60.07 (32.35)	65.82 (34.99)	.301
HADS depression	2.56 (3.09)	2.71 (3.62)	2.55 (3.07)	.680
HADS anxiety	4.21 (3.30)	4.59 (3.77)	4.19 (3.27)	.467
Time taken to read instructions (s)				
DET	11.17 (12.84)	12.69 (8.17)	11.10 (13.02)	.546
IDN	8.02 (21.82)	10.02 (11.17)	7.92 (22.20)	.546
OBK		6.79 (5.46)	7.15 (41.68)	.941
OCL	11.70 (35.11)	8.74 (6.43)	11.85 (35.91)	.575
Time taken to complete tests (s)				
DET	72.37 (19.90)	127.17 (54.31)	69.74 (11.15)	<.001
IDN	66.53 (17.18)	113.01 (53.86)	64.30 (7.95)	<.001
OBK	82.64 (19.96)	128.05 (45.05)	80.46 (14.75)	<.001
OCL	213.55 (24.88)	219.27 (37.79)	213.27 (24.07)	.044

Note: Chi-square was used to test differences between groups for categorical variables, and analysis of variance was used to test differences between groups for continuous variables; FDR corrections have been applied to all *P* values.
Abbreviations: DET, Detection test; FDR, false-discovery rate; HADS, Hospital Anxiety and Depression scale; IDN, Identification test; OCL, One Card Learning test; OBK, One-Back test.
[a]Failed error = participants did not meet pre-specified error criteria on at least two cognitive tests.

TABLE 3 Amount of time spent on instruction screen, and time taken to complete each cognitive test

	Overall (n = 1594) Mean (SD)	40 to 49 years (n = 321) Mean (SD)	50 to 59 years (n = 745) Mean (SD)	60 to 70 years (n = 528) Mean (SD)	P1	P2
Time spent on instruction screen (s)						
DET	11.17 (12.84)	9.73 (6.20)	11.47 (16.49)	11.64 (9.44)	.215	.090
IDN	8.02 (21.82)	6.50 (7.72)	8.74 (30.94)	7.93 (7.05)	.260	.352
OBK	7.13 (40.73)	4.93 (8.47)	5.73 (7.37)	10.45 (69.84)	.766	.092
OCL	11.70 (35.11)	10.36 (15.85)	11.30 (34.63)	13.09 (43.34)	.766	.339
Total CBB	38.03 (60.65)	31.51 (23.97)	37.24 (50.87)	43.12 (84.04)	.260	.035
N (%) who did not meet HCI learnability criteria for time spent on instruction screen						
DET	24 (1.5%)	3 (0.9%)	8 (1.1%)	13 (2.5%)	.903	.448
IDN	7 (0.4%)	2 (0.6%)	3 (0.4%)	2 (0.4%)	.903	.614
OBK	6 (0.4%)	1 (0.3%)	2 (0.3%)	3 (0.6%)	.903	.614
OCL	11 (0.7%)	4 (1.2%)	3 (0.4%)	4 (0.8%)	.472	.614
Time taken to complete cognitive test (s)						
DET	72.37 (19.90)	70.05 (16.42)	72.65 (20.40)	73.39 (21.01)	.064	.018
IDN	66.53 (17.18)	64.16 (10.16)	66.51 (17.50)	67.99 (19.79)	.064	.003
OBK	82.64 (19.96)	80.03 (15.98)	82.32 (20.09)	84.66 (21.69)	.085	.003
OCL	213.55 (24.88)	209.84 (22.02)	214.59 (27.93)	214.34 (21.55)	.010	.014
Total CBB	435.08 (58.78)	424.09 (44.97)	436.06 (60.87)	440.38 (62.29)	.010	.001
N (%) who did not meet HCI learnability criteria for time spent completing test						
DET	72 (4.5%)	8 (2.5%)	38 (5.1%)	26 (4.9%)	.128	.107
IDN	49 (3.1%)	4 (1.2%)	24 (3.2%)	21 (4.0%)	.128	.072
OBK	82 (5.1%)	11 (3.4%)	35 (4.7%)	36 (6.8%)	.349	.072
OCL	61 (3.8%)	10 (3.1%)	35 (4.7%)	16 (3.0%)	.317	.944

Note: FDR corrections have been applied to all P values; P^1 = 40 to 49 years versus 50 to 59 years; P^2 = 40 to 49 years versus 60 to 70 years; learnability cut-off was computed as 2 SDs above the mean time taken for individuals who passed error criteria, for test instruction: DET (37.1 s), IDN (52.3 s), OCL (83.7 s) and OBK (90.5 s); and for time spent completing each test: DET (92.0 s), IDN (80.2 s), OCL (261.4 s), and OBK (110.0 s).
Abbreviations: CBB, Cogstate Brief Battery; DET, Detection test; FDR, false-discovery rate; HADS, Hospital Anxiety and Depression scale; HCI, human computer interaction; IDN, Identification test; OCL, One Card Learning test; OBK, One-Back test.

(Table 2), we specified that HCI learnability was considered to have occurred if participants who exceeded 2 SDs from the average time taken to read the instruction screen or to complete tests based on the estimates of the group of participants who passed HCI error criteria.

We observed age-related increases in the amount of time taken to read test instructions, but only for the DET test (Table 3). Similarly, we observed age-related increases in the amount of time taken to complete all cognitive tests. Over 90% of participants met HCI learnability criteria for the test instruction screen, and for time taken to complete each cognitive test, with no differences across age groups (Table 3).

3.4 | Validity of unsupervised cognitive testing: relationship with test difficulty and age

For participants whose data satisfied HCI acceptability and HCI usability criteria, group mean (SD) performance on each cognitive test is summarized in Table 4 for the overall group and for each age-group. As predicted, speed of performance varied as a function of test difficulty. Paired samples t tests of the overall group performance indicated that average performance on the OCL test was significantly slower than on the OBK test, $t(1520) = 56.92, P < .001$, which in turn was significantly slower than the IDN test, $t(1520) = 101.22, P < .001$, which in turn was significantly slower than the DET test, $t(1520) = 75.006, P < .001$. Similarly, accuracy of performance on the DET test was significantly better than the IDN test, $t(1520) = 12.84, P < .001$, which in turn was significantly than the OBK test, $t(1520) = 5.62, P < .001$, and in turn was significantly between than the OCL test, $t(1520) = 87.55, P < .001$ (Table 4).

After adjusting for participants' years of education, we observed that speed of performance across most cognitive tests was significantly slower in older age groups (Table 4). When participants' age was considered continuously (and adjusting for years of education), speed of performance on all tests was significantly slower with older age; DET (β[standard error (SE)] = 0.19[0.03], $P < .001$), IDN (β[SE] = 0.23[0.03],

TABLE 4 Average speed and accuracy of performance for each cognitive test

	Overall (n = 1521)	40 to 49 years (n = 311)	50 to 59 years (n = 710)	60 to 70 years (n = 500)		
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P1	P2
DET speed	2.55 (0.10)	2.52 (0.09)	2.55 (0.10)	2.57 (0.09)	<.001	<.001
IDN speed	2.70 (0.06)	2.67 (0.07)	2.69 (0.05)	2.71 (0.07)	<.001	<.001
OBK speed	2.89 (0.09)	2.88 (0.09)	2.89 (0.11)	2.91 (0.09)	.055	<.001
OCL speed	3.01 (0.08)	2.99 (0.09)	3.01 (0.08)	3.01 (0.09)	<.001	.001
DET accuracy	1.48 (0.14)	1.48 (0.16)	1.48 (0.16)	1.48 (0.16)	.999	.999
IDN accuracy	1.42 (0.14)	1.42 (0.12)	1.42 (0.13)	1.41 (0.13)	.999	.463
OBK accuracy	1.39 (0.15)	1.40 (0.16)	1.39 (0.16)	1.38 (0.16)	.972	.267
OCL accuracy	1.00 (0.11)	1.01 (0.11)	1.01 (0.11)	0.99 (0.11)	.968	.056

Note: FDR corrections have been applied to all *P* values; all values have been adjusted for years of education; P^1 = 49 to 49 years versus 50 to 59 years; P^2 = 40 to 49 years versus 60 to 70 years.
Abbreviations: DET, Detection test; FDR, false-discovery rate; IDN, Identification test; OCL, One Card Learning test; OBK, One-Back test.

TABLE 5 Demographic characteristics of participants who completed tests in each testing environment

	Alone	With others (quiet)	With others (noisy)	
N	340	433	54	P
N (%) females	270 (79.4%)	323 (74.6%)	43 (79.6%)	.743
N (%) rural/regional location	91 (26.8%)	96 (22.2%)	10 (18.5%)	.685
N (%) white	258 (75.9%)	341 (78.8%)	43 (79.6%)	.745
Age	57.66 (6.29)	55.92 (6.84)	53.41 (6.78)	<.001
Education	12.00 (3.42)	11.96 (3.32)	11.43 (3.44)	.743
Annual income (self) ('000s)	62.99 (34.78)	66.38 (34.95)	69.34 (31.47)	.685
HADS depression	2.57 (3.27)	2.36 (3.00)	2.81 (3.37)	.743
HADS anxiety	3.98 (3.51)	3.95 (3.00)	4.35 (3.23)	.788
Total time to complete CBB (min)	7.15 (0.77)	7.11 (0.62)	6.95 (0.37)	.685
N (%) met overall HCI error criteria	321 (94.4%)	413 (95.4%)	51 (94.4%)	.819
N (%) met HCI learnability (DET)	322 (94.7%)	414 (95.6%)	52 (96.3%)	.787
N (%) met HCI learnability (IDN)	330 (97.1%)	418 (96.5%)	51 (94.4%)	.787
N (%) met HCI learnability (OBK)	317 (93.2%)	413 (95.4%)	52 (96.3%)	.720
N (%) met HCI learnability (OCL)	322 (94.7%)	414 (95.6%)	54 (100%)	.720

Note: FDR corrections have been applied to all *P* values.
Abbreviations: CBB, Cogstate Brief Battery; DET, Detection test; FDR, false-discovery rate; HADS, Hospital Anxiety and Depression scale; HCI, Human computer interaction; IDN, Identification test; OBK, One-Back test; OCL, One Card Learning test.

P < .001), OBK (β[SE] = 0.13[0.03], *P* < .001), OCL (β[SE] = 0.12[0.03], *P* < .001). Accuracy of performance did not differ significantly between age groups on any tests (Table 4). However, when age was considered continuously, and after accounting for education, increasing age was significantly associated with poorer accuracy for the OBK (β[SE] = −0.06[0.0303], *P* = .045) and OCL (β[SE] = −0.07[0.0303], *P* = .021) tests, although these associations were, by convention, very small in magnitude.

3.5 | Effect of testing environment, and first-degree family history of dementia on cognitive performance

A subsample of participants (n = 827; 62%) who completed cognitive tests also completed a survey of their usual testing environment. Table 5 provides the demographic characteristics of this subset of participants. Most participants completed cognitive tests at home/work

TABLE 6 Effect of testing environment, and first-degree family history of dementia, on the primary outcome of each cognitive test

Effect of testing environment								Alone (n = 340)	With others (quiet) (n = 433)	With others (noisy) (n = 54)
	Age		Education		Group					
	(df) F	P	(df) F	P	(df) F	P		Mean (SD)	Mean (SD)	Mean (SD)
DET speed	(1760) 37.01	<.001	(1760) 5.64	.018	(2760) 0.37	.691		2.54 (0.005)	2.55 (0.004)	2.55 (0.013)
IDN speed	(1760) 54.62	<.001	(1760) 2.73	.100	(2760) 0.001	.999		2.69 (0.004)	2.70 (0.003)	2.70 (0.009)
OBK accuracy	(1760) 2.83	.093	(1760) 10.57	.001	(2760) 0.10	.903		1.39 (0.009)	1.39 (0.007)	1.40 (0.021)
OCL accuracy	(1760) 4.33	.038	(1760) 20.49	<.001	(2760) 0.85	.427		1.01 (0.006)	1.00 (0.005)	1.02 (0.016)
Effect of first-degree family history of dementia								1° family hx (n = 674)	Without 1° family hx (n = 704)	
	Age		Education		Group					
	(df) F	P	(df) F	P	(df) F	P		Mean (SD)	Mean (SD)	
DET speed	(11,358) 10.84	<.001	(11,358) 45.90	<.001	(11,358) 10.21	.001		2.56 (0.104)	2.54 (0.106)	
IDN speed	(11,358) 75.25	<.001	(11,358) 9.76	.002	(11,358) 9.26	.002		2.70 (0.052)	2.69 (0.053)	
OBK accuracy	(11,358) 5.52	.019	(11,358) 8.81	.003	(11,358) 0.17	.684		1.39 (0.156)	1.39 (0.159)	
OCL accuracy	(11,358) 7.08	.008	(11,358) 24.43	<.001	(11,358) 0.02	.901		1.00 (0.104)	1.00 (0.106)	

Note: FDR corrections have been applied to all *P* values.
Abbreviations: DET, Detection test; FDR, false-discovery rate; IDN, Identification test; OBK, One-Back test; OCL, One Card Learning test.

without other people present (41.1%) or at home/work where other people are present but are quiet (52.4%). Participants in the "with others (noisy)" group were significantly younger than those in the "with others (quiet)" group (*P* = .011, *d* = 0.37), who were in turn younger than participants in the "alone" group (*P* < .001, *d* = 0.26). Error and learnability did not vary significantly across each of the testing environments (Table 5).

The effect of testing environment on the primary outcome measure of each cognitive test, after accounting for the effects of age and education, are summarized in Table 6. Age- and education-adjusted performance mean (SD) of each testing environment group are also provided in Table 6. We observed no overall effect of testing environment on any cognitive outcome measure. However, post-hoc comparison suggests that participants in the "with others (quiet)" group performed significantly slower than those in the "alone" group on the IDN test, but the magnitude of difference was very small, *d*(95% confidence interval [CI]) = 0.15 (0.01, 0.29). No other between-group comparisons were significant, all *P*'s > .35, all *d*'s < 0.10.

Table 6 also summarizes the effect of first-degree family history of dementia on cognitive function, accounting for the effects of age and education. We observed an overall effect of first-degree family history of dementia only on the DET and IDN tests, and the magnitude of difference between groups was very small and identical for both tests, *d*(95% CI) = 0.19 (0.08, 0.30; Table 6). No other between-group comparisons were significant, all *P*'s > .900.

4 | DISCUSSION

The aim of this study was to determine the acceptability and usability of unsupervised cognitive testing. With those characteristics established, we also aimed to determine the validity of unsupervised cognitive testing in middle-aged adults enrolled in a large online study of AD risk. A large group of adults from metropolitan, regional, and rural Australia enrolled in the HBP completed an online battery of computerized cognitive tests. In this group, we showed that with appropriate design, a battery of unsupervised cognitive tests showed high acceptability and usability. Low rates of missing data were observed (1.9%), which occurred only for the first and second tests in the battery. This pattern suggests that missing data are more likely to occur in earlier tasks, when participants are less familiar with the battery, and indicates that as participants became more comfortable with the battery as the tests progressed they were able to complete the battery successfully. The current study sample also demonstrated high levels of understanding and adherence with the rules and requirements of each test, as measured by the extent to which performance on each test fell within expected normative limits. Limits of accuracy of performance were obtained from normative data for the same tests administered to healthy adults in supervised settings.[20,32] Using these criteria, we observed that error rates were consistently low across all tests (~10%), and when considering all tests, <5% of participants failed error criteria on two or more tests. Individuals in the HBP sample required ~10 seconds to read the instructions for each test, and except for the first test presented (DET), the average amount of time taken to read these test instructions did not increase with age. Participants also demonstrated high rates of learnability across all test instructions (~99%) and cognitive tests (~95%), and this did not vary with age (Table 3). As expected, the time required to complete each test varied with test difficulty and length, and this time requirement increased slightly with age. Together, these data suggest that the unsupervised cognitive tests applied in the HBP are acceptable and usable in the

samples sought. They also provide a guide for the development and application of automated analyses to identify low quality test data and differentiate this from data that reflects true cognitive performance.

The second hypothesis, that expected test difficulty–related, and age-related decreases in performance would be demonstrated was also supported. First, we observed that speed and accuracy of performance worsened as test complexity increased across the simple reaction time (DET), choice reaction time (IDN), working memory (OBK), and the pattern separation learning (OCL) tests. This influence of test difficulty on performance is well known in the psychological literature and has been observed many times in the context of supervised assessments using these same cognitive tests.[18–20] The presence of these relationships in the HBP sample, therefore, provides strong support that these tests are measuring those aspects of cognition they were designed to measure even when given in an unsupervised context. The second aspect of validity demonstrated was that older age was associated with slower reaction time on all tests, and with poorer accuracy of performance on the OBK and OCL tests. Age-related decline in cognition, even in the absence of neurodegenerative disease, has been well described,[33–36] including in unsupervised settings.[25] While these effects are small, they do indicate that data arising from the application of these tests in unsupervised settings did retain their validity.

Finally, in a subset of participants, we sought to explore the impact of participants' self-reported testing environment on cognitive performance. Of the 1594 participants who completed cognitive testing, ~65% (n = 827) completed a survey of their usual testing environment (defined as either: alone; with others [quiet]; with others [noisy]). When accounting for the effects of age, no general influence of testing environment was found for any of the cognitive outcome measures in this study, except for a small (d = 0.14) effect of reduced speed of performance in those who completed testing in the "alone" group compared to "with others [quiet]." These findings contrast with those of the existing literature, which suggests performance on unsupervised neuropsychological tests may be particularly sensitive to test environment.[13] For example, background music and noise has been reported to impair performance on cognitive tests.[37,38] When considered along with the observation that aspects of error and learnability did not vary as a result of testing environment, our findings do provide further support for the use of the CBB in unsupervised cognitive testing.[18] However, future research comparing the performance of the same individual across different testing environments in both unsupervised and supervised contexts is required to further support this conclusion.

An important limitation to our study is that it is cross-sectional in design, and as such, information about participant retention and attrition were not available. Additionally, as this study was administered completely online, we did not obtain any data related to the supervised (in-person) administration of the CBB. A direct comparison of unsupervised and supervised CBB testing may have strengthened the impact of the findings of this study by providing further information regarding the acceptability and validity of unsupervised assessment. We also acknowledge that our comparison of performance across varied testing environments may be limited by both power (large differences in sample size between the "noisy" [n = 54], "with others [quiet]" [n = 433]

and "alone" [n = 340] groups) and that we did not experimentally determine performance of the same individual in different testing environments. It should, however, be noted that the testing environment survey was designed to reflect the flexibility of the HBP testing schedule (ie, participants were not required to complete all surveys in a single sitting). Consequently, there may be some disconnect between the actual environment in which the CBB was completed and where the participant completed the majority of their assessments. Finally, it is important to note that the HBP does not randomly sample the population. Family history was used as a proxy for AD or dementia risk, and while having a family history is not a very strong predictor of AD, this strategy has resulted in higher proportions of apolipoprotein E ε4 carriers in the study sample compared to the general population (35% vs 18%).[23]

While we have established the internal validity of these tests, we need to now determine the external validity by using an established criterion for abnormality. For example, future studies will need to examine the acceptability, usability, and validity of performance of individuals with clinically diagnosed mild cognitive impairment (MCI). Recently, poorer performance across all CBB tests was observed in self-reported MCI patients and self-reported AD patients compared to self-reported healthy controls who completed unsupervised testing.[25] While promising, additional studies are needed to determine the utility of unsupervised cognitive tests in detecting cognitive impairment in clinically confirmed MCI and AD patients.

These caveats notwithstanding, the data in this study show how the HCI approach can provide a suitable foundation for development and refinement of unsupervised cognitive tests. With acceptability and usability, we showed that psychometric characteristics of cognitive test data generated in unsupervised contexts can be challenged using conventional psychometric approaches. We also examined the validity of unsupervised cognitive tests by determining the extent to which the expected effects of age and test difficulty manifest in performance. However, other psychometric approaches such as examination of factor structure, differential item function, as well as the more crucial characteristics such as sensitivity to AD related risk factors or to AD related cognitive change, can also be applied to understand data generated in unsupervised settings. With these factors considered, the data collected in this study do appear to retain similar psychometric characteristics as those collected from supervised testing of the same tests. As such, these results support the acceptability, usability, and validity of the CBB in the unsupervised assessment of cognition in individuals at risk of dementia. The approach used here can also be applied to other cognitive tests and surveys administered via online or web-based platforms for the unsupervised assessment of individuals with self-reported cognitive impairment[25] or those who are at risk of dementia.[23]

ACKNOWLEDGMENTS

CONFLICTS OF INTEREST

S Perin, RF Buckley, MP Pase, N Yassi, and YY Lim report no disclosures. A Schembri and P Maruff are full-time employees of Cogstate Ltd.

REFERENCES

1. Jack CR, Jr., Knopman DS, Jagust WJ, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol*. 2010;9:119-128.
2. Langbaum JB, Fleisher AS, Chen K, et al. Ushering in the study and treatment of preclinical Alzheimer disease. *Nat Rev Neurol*. 2013;9(7):371-381.
3. Sperling RA, Aisen PS, Beckett LA, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7:280-292.
4. Insel PS, Weiner M, Mackin RS, et al. Determining clinically meaningful decline in preclinical Alzheimer disease. *Neurology*. 2019;93:322-333.
5. Buckley RF, Mormino EC, Amariglio RE, et al. Sex, amyloid, and APOE ε4 and risk of cognitive decline in preclinical Alzheimer's disease: Findings from three well-characterized cohorts. *Alzheimers Dement*. 2018;14:1193-1203.
6. Ellis KA, Bush AI, Darby D, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr*. 2009;21:672-687.
7. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology*. 2010;74:201-209.
8. Bauer RM, Iverson GL, Cernich AN, Binder LM, Ruff RM, Naugle RI. Computerized neuropsychological assessment devices: joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Arch Clin Neuropsychol*. 2012;27:362-373.
9. Feenstra HEM, Vermeulen IE, Murre JMJ, Schagen SB. Online cognition: factors facilitating reliable online neuropsychological test results. *Clin Neuropsychol*. 2017;31:59-84.
10. Rolbillard JM, Illes J, Arcand M, et al. Scientific and ethical features of English-language online tests for Alzheimer's disease. *Alzheimers Dement*. 2015;1:281-288.
11. Weiner MW, Nosheny R, Camacho M, et al. The Brain Health Registry: an internet-based platform for recruitment, assessment, and longitudinal monitoring of participants for neuroscience studies. *Alzheimers Dement*. 2018;14(8):1063-1076.
12. Feenstra HEM, Vermeulen IE, Murre JMJ, Schagen SB. Online cognition: factors facilitating reliable online neuropsychological test results. *Clin Neuropsychol*. 2016;31:59-84.
13. Crump MJ, McDonnell JV, Gureckis TM. Evaluating Amazon's mechanical turk as a tool for experimental behavioral research. *PLoS One*. 2013;8(3):e57410.
14. Neilsen AS, Wilson RL. Combining e-mental health intervention development with human computer interaction (HCI) design to enhance technology-facilitated recovery for people with depression and/or anxiety conditions: an integrative literature review. *Int J Ment Health Nurs*. 2019;28:22-39.
15. Lawrence DO, Ashleigh MJ. Impact of human computer interaction (HCI) on users in higher educational system: Southampton University as a case study. *Int J Manag Technol*. 2019;6:1-12.
16. Masooda M, Thigambaram M. *The Usability of Mobile Applications for Pre-Schoolers*. Athens, Greece: 7th World Conference on Educational Sciences; 2015:1818-1826.
17. Shackel B. Usability—context, framework, definition, design and evaluation. *Interact Comput*. 2009;21:339-346.
18. Cromer JA, Harel BT, Yu K, et al. Comparison of cognitive performance on the Cogstate Brief Battery when taken in-clinic, in-group, and unsupervised. *Clin Neuropsychol*. 2015;29:542-558.
19. Darby DG, Fredrickson J, Pietrzak RH, Maruff P, Woodward M, Brodtmann A. Reliability and usability of an internet-based computerized cognitive testing battery in community-dwelling older people. *Comput Hum Behav*. 2014;30:199-205.
20. Fredrickson J, Maruff P, Woodward M, et al. Evaluation of the usability of a brief computerized cognitive screening test in older people for epidemiological studies. *Neuroepidemiology*. 2010;34:65-75.
21. Stiles-Shields C, Montague E, Lattie EG, Schueller SM, Kwasny MJ, Mohr DC. Exploring user learnability and learning performance in an app for depression: usability study. *JMIR Hum Factors*. 2017;4:e18.
22. Holzinger A. Usability engineering methods for software developers. *Commun ACM*. 2005;48:71-74.
23. Lim YY, Yassi N, Bransby L, Properzi MJ, Buckley R. The Healthy Brain Project: an online platform for the recruitment, assessment, and monitoring of middle-aged adults at risk of developing Alzheimer's disease. *J Alzheimers Dis*. 2019;68:1211-1228.
24. Harrison R, Flood D, Duce D. Usability of mobile applications: literature review and rationale for a new usability model. *J Interact Sci*. 2013;1:1.
25. Mackin RS, Insel PS, Truran D, et al. Unsupervised online neuropsychological test performance for individuals with mild cognitive impairment and dementia: results from the Brain Health Registry. *Alzheimers Dement*. 2018;10:573-582.
26. Lim YY, Ellis KA, Harrington K, et al. Use of the CogState Brief Battery in the assessment of Alzheimer's disease related cognitive impairment in the Australian Imaging, Biomarkers and Lifestyle (AIBL) study. *J Clin Exp Neuropsychol*. 2012;34:345-358.
27. Sternberg S. Modular processes in mind and brain. *J Cogn Neuropsychol*. 2011;28:156-208.
28. Collie A, Maruff P, Darby DG, McStephen M. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. *J Int Neuropsychol Soc*. 2003;9:419-428.
29. Maruff P, Lim YY, Darby D, et al. Clinical utility of the Cogstate Brief Battery in identifying cognitive impairment in mild cognitive impairment and Alzheimer's disease. *BMC Pharmacol Toxicol*. 2013;1:1-11.
30. Lim YY, Villemagne VL, Laws SM, et al. Performance on the Cogstate Brief Battery is related to amyloid levels and hippocampal volume in very mild Dementia. *J Mol Neurosci*. 2016;60:362-370.
31. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29:1165-1188.
32. Moriarity JM, Pietrzak RH, Kutcher JS, Clausen MH, McAward K, Darby DG. Unrecognised ringside concussive injury in amateur boxers. *Br J Sports Med*. 2012;46(14):1011-1015.
33. Boyle PA, Wilson RS, Yu L, et al. Much of late life cognitive decline is not due to common neurodegenerative pathologies. *Ann Neurol*. 2013;74(3):478-489. https://doi.org/10.1002/ana.23964.
34. Drag LL, Bieliauskas LA. Contemporary review 2009: cognitive aging. *J Geriatr Psychiatry Neurol*. 2009;23:75-93.

35. Lipnicki DM, Sachdev PS, Crawford J, et al. Risk factors for late-life cognitive decline and variation with age and sex in the Sydney Memory and Ageing Study. *PLoS One.* 2013;8:e65841.

36. Manard M, Carabin D, Jaspar M, Collette F. Age-related decline in cognitive control: the role of fluid intelligence and processing speed. *BMC Neurosci.* 2014;15:7.

37. Cassidy G, MacDonald R. The effects of background music and background noise on the task performance of introverts and extraverts. *Psychol Music.* 2007;35:517-537.

38. Dobbs S, Furnham A, McCelland A. The effect of background music and noise on the cognitive test performance of introverts and extraverts. *Appl Cogn Psychol.* 2011;25:307-313.