CORTEX 145 (2021) 264–272

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

Automated analysis of propositional idea density in older adults

Edward Hill [a,*], *Jane Alty* [a], *Larissa Bartlett* [a], *Lyn Goldberg* [a], *Mira Park* [b], *Soonja Yeom* [b] *and James Vickers* [a]

[a] *Wicking Dementia Research and Education Centre, University of Tasmania, Hobart, Tasmania, Australia*
[b] *Information and Communication Technology, University of Tasmania, Hobart, Tasmania, Australia*

ARTICLE INFO

Article history:
Received 22 June 2021
Reviewed 26 July 2021
Revised 29 July 2021
Accepted 28 September 2021
Action editor Peter Garrard
Published online 23 October 2021

Keywords:
Language
Dementia
Propositional idea density
Cognitive function

ABSTRACT

Previous research suggests oral and written language can act as barometers of an individual's cognitive function, potentially providing a screening tool for the earliest stages of Alzheimer's disease (AD) and other forms of dementia. Idea density is a measure of the rate at which ideas, or elementary predications, are expressed and may provide an ideal measure for early detection of deficits in language. Previous research has shown that when no restrictions are set on the topic of the idea, a decrease in propositional idea density (PID) is associated with an increased risk of developing AD. However, this has been limited by moderate sample sizes and manual transcribing. Technological advancement has enabled the automated calculation of PID from tools such as the Computerized Propositional Idea Density Rater (CPIDR). We delivered an online autobiographical writing task to older adult Australians from ISLAND (Island Study Linking Ageing and Neurodegenerative Disease). Linear regression models were fitted in R. We analysed text files (range 10–1180 words) using CPIDRv5 provided by 3316 (n = 853 males [25.7%], n = 2463 females [74.3%]) ISLAND participants. Over 358,957 words written in 3316 written autobiographical responses were analysed. Mean PID was higher in females (53.5 [±3.69]) than males (52.6 [±4.50]). Both advancing age and being male were significantly associated with a decrease in PID ($p < .001$). Automated methods of language analysis hold great promise for the early detection of subtle deficits in language capacity. Although our effect sizes were small, PID may be a sensitive measure of deficits in language in ageing individuals and is able to be collected at scale using online methods of data capture.

* *Corresponding author*. Wicking Dementia Research and Education Centre, University of Tasmania, Medical Sciences Precinct 1, 17 Liverpool St, Hobart, Tasmania, 7000, Australia.
E-mail addresses: e.hill@utas.edu.au (E. Hill), jane.alty@utas.edu.au (J. Alty), larissa.bartlett@utas.edu.au (L. Bartlett), lyn.goldberg@utas.edu.au (L. Goldberg), mira.park@utas.edu.au (M. Park), soonja.yeom@utas.edu.au (S. Yeom), james.vickers@utas.edu.au (J. Vickers).
https://doi.org/10.1016/j.cortex.2021.09.018

1. Introduction

Language is a window into the brain, and has been proposed as an early detection tool for neurodegenerative disease, such as Alzheimer's disease (AD) (Christensen et al., 2008; Laguarta et al., 2020). An individual's ability to express language, whether through spoken word or written text, may provide insight about their underlying cognitive processes. In AD, language deficits are noticeable in the earliest stages of the disease continuum (Morris, 1996). For this reason, oral and written language has been described as barometer of cognitive function and a valid indicator of an individual's 'cognitive reserve' (Christensen et al., 2008), or their neuroresilience to age related biological changes, potentially providing an opportunity to deploy a simple screening test that detects subtle deficits.

Without pharmacological agents demonstrating efficacy to halt the neurodegenerative cascade, early detection and management of AD risk is of great importance. With an ageing population and increasing incidence of ageing-related conditions such as AD and other forms of dementia, there remains a great need for non-invasive, cost-effective and sensitive tools that can aid in early detection and subsequent intervention. AD is a disease of aging, yet not a normal part of aging, with the prevalence increasing with age; 3% at age 65–74, 17% at age 75–84 and 32% for people aged 85 and over (Hebert et al., 2013). As AD develops approximately 20 years before symptoms arise, early detection would allow for targeted interventions that could improve trajectories before cognitive symptoms emerge, enabling individuals to live better for longer. Further, the rapid promulgation of online methods of audio, video and text data collection has enabled scalable and accurate measurement of cognitive processes via the internet. Large scale epidemiological cohort studies can collect data with sufficient power for detecting subtle language and cognitive deficits in the general population. Methods for the automated detection of language deficits in cognitive decline are gaining traction (Komeili et al., 2019; Laguarta et al., 2020), with online administration removing geographical, social and economic barriers to participation.

Idea density may provide an ideal measure for early detection of language deficits. Idea density measures the rate at which ideas, or elementary predications, are expressed (Turner & Greene, 1997). Idea density can be broken down into propositional idea density (PID), which does not set restrictions on the topic of the idea, or semantic idea density, which relies on a pre-defined set of information content units. PID can be calculated from spoken or written words and lower PID is associated with an increased risk of developing AD (Engelman et al., 2010) and AD pathology (Snowdon et al., 1996). For this reason, PID may be a candidate measure of cognitive function as measured through language and a valid predictor of the onset and progression of AD and other forms of dementia (Kemper et al., 1993; Lyons et al., 1994; Snowdon et al., 1996). Historical methods of PID calculation relied on time-consuming manual coding, which has limited the ability to process data from large samples. However, recent advances in technology have enabled fast, accurate and scalable measurements of PID from written text. Automatic PID calculation is now possible using the Computerized Propositional Idea Density Rater (CPIDR) (Brown et al., 2008) and has been widely used in language analysis (Clarke et al., 2020; Ferguson et al., 2014; Kim et al., 2019). CPIDR is a computer program that determines the PID of English text automatically against part-of-speech tags. Based on the idea that propositions correspond roughly to verbs, adjectives, adverbs, prepositions and conjunctions, CPIDR tags speech and applies numerous rules to generate PID for a given text. CPIDR has been tested and validated against human raters (Brown et al., 2008).

Previous research has demonstrated the predictive utility of PID in ageing and dementia research yet has failed to present normative ranges of typical PID over time. Several studies have shown decreases in PID with age. Participants from the Nun Cohort (n = 139) demonstrated declines in PID with age, older adults (mean age 78 years) demonstrating a mean (±standard deviation [SD]) PID of 36 (±11) – 55 (±5), while their young adult (mean age 22 years) counterparts had a PID of 53 (±6) – 72(±11) (Kemper, Greiner, et al., 2001). Participants wrote their first autobiography at a mean age of 22 years, the second at 83 years. Kemper, Greiner, et al. (2001) reported declines in PID (.3 units per year) at a similar rate across a variety of age gaps, indicating the decline did not depart from linearity and appeared uniform throughout the lifespan. Kemper and Sumner (2001) investigated PID in younger (age range 18–28 years) and older (age range 63–88) adults, finding PID was significantly lower in the older adults (n = 200). However, this study was limited by a lack of participants in the middle adult age group, thus failing to draw conclusions over their age-related trajectories. Further, Kemper, Marquis, and Thompson (2001) fit linear regression models to describe associations between age and PID in a longitudinal cohort study (15 years) made up of healthy controls and participants with dementia. Both time and age were significantly associated with PID, however this study was limited by a smaller sample size (n = 30) and attrition was higher in the group with dementia.

Together with Kemper, Greiner, et al. (2001), similar studies have also demonstrated significant declines of PID with age (Ferguson et al., 2014; Kemper, Marquis, & Thompson, 2001; Kemper & Sumner, 2001), whilst only one study has reported a significant association with educational attainment (Kemper, Greiner, et al., 2001). As education is widely regarded to build an individual's cognitive reserve and in previous research PID has not demonstrated consistent associations with education, PID may represent an ideal marker of cognitive reserve that is unrelated to education, thus reflecting other aspects of cognitive stimulation.

Published findings on PID in epidemiological cohorts tend to be of small to moderate sample sizes (n ≤ 200 participants) (Engelman et al., 2010; Farias et al., 2012; Kemper, Greiner, et al., 2001; Kemper & Sumner, 2001; Riley et al., 2005; Snowdon et al., 1996) and/or only female participants (Ferguson et al., 2014; Kemper, Greiner, et al., 2001; Snowdon et al., 1996). Given sex differences in AD symptomatology, pathology and cognition (Ferretti et al., 2018), it is important to consider the effect of sex on PID, including both males and females in analysis. Whilst the technological infrastructure is available, there has been little examination of automated PID in a large-scale epidemiological

cohort study. Such research would provide age-normative data of a sex-stratified and demographically diverse cohort, whilst proving relevant to the growing field of early detection and screening test development in AD research. In this cross-sectional study, we aimed to measure PID in over 3,000 males and females aged 50 and over, drawn from the Tasmanian longitudinal dementia risk reduction initiative, the ISLAND Project: Island Project Linking Ageing and Neurodegenerative Disease. We hypothesized that higher age, but not higher education, would be associated with a reduction in PID.

2. Methods

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria and whether they were established prior to data analysis, all manipulations, and all measures in the current study.

2.1. Study population

We invited participants from the ISLAND research cohort to complete the protocol. ISLAND is a world-first initiative to target dementia incidence at the population level by addressing dementia risk through behaviours, knowledge and self-efficacy. ISLAND aims to reduce dementia incidence through nested interventions that target known modifiable risk factors (Livingston et al., 2020). A comprehensive protocol is available elsewhere (Bartlett et al., 2021). Briefly, participating in ISLAND research is open to all adults aged 50 years or older, residing in the island state of Tasmania, Australia. Participants complete surveys on knowledge, attitudes and behaviours related to modifiable dementia risk factors. These data, together with intervention engagement indices, cognitive functioning and blood-based biomarkers are used to track and investigate changes over time. In the current study, all data was collected online, via the participant's own ISLAND Home Portal, an online interface between researchers and participants. All ISLAND participants were aged 50 years or over at time of recruitment. Age (in years) was calculated from their entered date of birth. Sex (male or female) was provided at study enrolment. Education (in years) was collected, as well as participants' highest level of education, from primary (elementary) school to a higher university degree (Honours, Graduate Diploma, Masters or PhD). For the purposes of analysis, we collapsed primary (kindergarten to year 6) and secondary (years 7–12) education into a single variable: primary/secondary education (in years). Background and health surveys asked about family history of dementia and the presence of other medical conditions diagnosed during participants' lifetime, such as dementia, heart disease, psychological disorders and cancer.

2.2. Autobiographical task

Participants were instructed to provide a short autobiography, similar to the protocols administered in previous studies (Brown et al., 2008; Kemper et al., 2001a, 2001b). Via their ISLAND Home Portal, the following instructions and probes were provided, with an open text box for data entry: *"Please take a few minutes to write a little bit about your history. We would like you to write between 120 and 200 words. That's about 10-15 sentences. If it's helpful you can choose one of the following prompts to get you started: Tell us about the town city place you grew up in. Describe your favourite holiday memory e.g., where you went, what you did. Tell us about someone you admire and why. Describe an unexpected event that happened to you and how it impacted your life."*

2.3. Propositional idea density

Autobiographical reflection data was split into individual by-participant text files and analysed in Computerized Propositional Idea Density Rater (CPIDR), accessed from the University of Georgia (http://ai1.ai.uga.edu/caspr/) under academic non-commercial access. CPIDR v5.1 is proprietary software belonging to the University of Georgia Research Foundation. CPIDR v5.1 uses a rule set to tag text files, providing idea (proposition count), word count, PID, the 95% confidence interval (CI) and an identifying string (first 37 characters of the text file). For example, the sentence: "*And every night I dream of the sea, they say home is where you find it*" provides 16 total words and 7 propositions/ideas ("*and*", "*every*", "*dream*", "*of*", "*say*", "*where*", "*find*"), resulting in a PID of 43.8 (95% CI 19.4–68.1). Further details are available in Brown et al. (2008).

2.4. Statistical analysis

This is a cross-sectional investigation from an ongoing longitudinal epidemiological cohort study. We excluded 17 autobiographical files as they provided less than the required minimum of 10 words, a lower limit proposed by Ferguson et al. (2014) to indicate complete sentence construction. PID was expressed as a value out of 100, whereby a value of 100 would indicate 100 ideas in 100 words, a zero indicating 0 ideas in 100 words. Linear regression models were fitted using the *lm* package in the base R statistical environment version 1.4, with age (in years), sex, education (in years), university (in years) included in our models as covariates for adjustment. Model 1 adjusted for age. Model 2 adjusted for age and sex. Model 3 adjusted for age, sex and education. Age (in years) was mean-centered and standardized by dividing each mean value by the SD. Unless otherwise stated, means, SDs and 95% CIs are reported. Correlations were analysed using Pearson coefficient.

2.5. Ethics

ISLAND has been approved by University of Tasmania's Health and Medical Human Research Ethics Committee (HREC H001864). This study was carried out in accordance with the National Health and Medical Research Council's National Statement on Ethical Conduct in Human Research (2018 update). All participants completed informed consent prior to any procedure in accordance with the Declaration of Helsinki. No part of this study procedures or analyses were pre-registered in a time-stamped, independent, institutional registry.

3. Results

3.1. Characteristics of the cohort

A total of 3,333 ISLAND participants completed the written reflection task, writing a single autobiographical passage of text. Of these, 3,316 provided complete demographic data and 3,190 also completed the background and health surveys. Characteristics of the cohort are provided in Table 1. Participants demonstrated an average age of 63.5 years (±7.80) with 11.5 (±1.41) years of primary/secondary education and 5.36 (±2.72) years of university education. The modal highest educational level reported was a higher university degree, with 1057 (31.9%) of participants reporting completing an Honours, Graduate Diploma, Masters or PhD. A total of 2463 (74.3%) participants reported their biological sex as female and displayed similar levels of primary/secondary and university education, although were slightly younger than their male counterparts (62.9 [±7.58] *vs.* 65.1 [±8.19] vs.).

Family history of dementia and medical comorbidities for ISLAND participants included in the current study are presented in Table 2. Approximately half of the participants included in this analysis reported a family history of dementia (n = 1484, 44.8%). ISLAND participants' most common comorbidities reported were psychological disorder (n = 693, 20.9%), hearing impairment (n = 518, 15.6%), cancer (n = 448, 13.5%) and heart disease (n = 266, 8.0%).

3.2. Propositional idea density

ISLAND participants provided autobiographical texts on average 203 (±104) words in length (range 10–1180 words), expressing 108 (±55.7) ideas resulting in an average PID of 53.3 (±3.94). Compared to males, females averaged slightly higher ideas (102 [±52.9] *vs.* 110 [±56.6]), words (195 [±105] *vs* 205 [±104]) and PID (52.6 [±4.50] *vs* 53.5 [±3.69]). PID stratified by categorical age groups is provided in Supplementary Table 1.

Results from linear regression analyses are presented in Table 3. In Model 1 that included only age, linear regression analysis demonstrated a significant decrease in PID with increasing age ($\beta = -.40$ [CI -.53 to −.26], $R^2 = .01$, $p < .001$) (Fig. 1). After the addition of sex as a covariate in Model 2, both increasing age ($\beta = -.35$ [CI -.48 to −.22], $R^2 = .02$, $p < .001$) and being male ($\beta = -.84$ [CI -1.14 to −.54], $R^2 = .02$, $p < .001$) were associated with a significant decrease in PID (Fig. 2). Males had .84 (CI -1.14 to −.54) lower PID than females, while for all participants each year of age increase was associated with a

Table 1 – Demographic and summary statistics for all ISLAND participants (n = 3316) included in subsequent analysis. Expressed are means and standard deviations (SD) unless otherwise indicated. Omitted are either not applicable (NA) or missing data.

	Male (N = 853)	Female (N = 2463)	Overall (N = 3316)
Age (in years)			
Mean (SD)	65.1 (8.20)	62.9 (7.58)	63.5 (7.80)
Median [Min, Max]	65.0 [50.0, 91.0]	63.0 [50.0, 94.0]	63.0 [50.0, 94.0]
Education (in years)			
Mean (SD)	11.5 (1.47)	11.5 (1.39)	11.5 (1.41)
Median [Min, Max]	12.0 [3.00, 20.0]	12.0 [0, 20.0]	12.0 [0, 20.0]
University Education (in years)			
Mean (SD)	5.62 (2.78)	5.27 (2.69)	5.36 (2.72)
Median [Min, Max]	5.00 [1.00, 20.0]	5.00 [0, 20.0]	5.00 [0, 20.0]
Highest Education Completed			
Primary School	1 (.1%)	2 (.1%)	3 (.1%)
High School	91 (10.7%)	310 (12.6%)	401 (12.1%)
Certificate or Apprenticeship (including Cert 2, 3 or 4)	103 (12.1%)	224 (9.1%)	327 (9.9%)
Diploma/Associate Degree	127 (14.9%)	432 (17.5%)	559 (16.9%)
Bachelor's Degree	185 (21.7%)	526 (21.4%)	711 (21.4%)
Higher University degree (Honours, Graduate Diploma, Masters or PhD)	283 (33.2%)	774 (31.4%)	1057 (31.9%)
Other	15 (1.8%)	74 (3.0%)	89 (2.7%)
Marital Status			
Single	52 (6.1%)	197 (8.0%)	249 (7.5%)
Defacto	87 (10.2%)	239 (9.7%)	326 (9.8%)
Married	573 (67.2%)	1320 (53.6%)	1893 (57.1%)
Separated or divorced	60 (7.0%)	366 (14.9%)	426 (12.8%)
Widowed	22 (2.6%)	181 (7.3%)	203 (6.1%)
Prefer not to say	0 (0%)	4 (.2%)	4 (.1%)
Other	2 (.2%)	21 (.9%)	23 (.7%)
Ideas			
Mean (SD)	103 (52.8)	110 (56.6)	108 (55.7)
Median [Min, Max]	92.0 [3.00, 405]	98.0 [5.00, 544]	97.0 [3.00, 544]
Words			
Mean (SD)	196 (105)	205 (104)	203 (104)
Median [Min, Max]	177 [10.0, 1180]	183 [13.0, 1010]	182 [10.0, 1180]
Propositional Idea Density (PID)			
Mean (SD)	52.6 (4.34)	53.5 (3.69)	53.3 (3.89)
Median [Min, Max]	52.7 [30.0, 69.0]	53.7 [31.6, 67.9]	53.5 [30.0, 69.0]

Table 2 – Reported prevalence of family history of dementia and medical comorbidities in participants (n = 3316) of the ISLAND Project, stratified by sex (male/female). Omitted are N/A or missing data.

	Male (N = 853)	Female (N = 2463)	Overall (N = 3316)
Family history of dementia			
Yes	354 (41.5%)	1130 (45.9%)	1484 (44.8%)
No	453 (53.1%)	1218 (49.5%)	1671 (50.4%)
Diagnosis: Dementia			
Yes	2 (.2%)	0 (0%)	2 (.1%)
No	813 (95.3%)	2353 (95.5%)	3166 (95.5%)
Diagnosis: Memory impairment			
Yes	21 (2.5%)	12 (.5%)	33 (1.0%)
No	789 (92.5%)	2339 (95.0%)	3128 (94.3%)
Diagnosis: Heart disease			
Yes	122 (14.3%)	144 (5.8%)	266 (8.0%)
No	687 (80.5%)	2202 (89.4%)	2889 (87.1%)
Diagnosis: Cancer			
Yes	139 (16.3%)	309 (12.5%)	448 (13.5%)
No	667 (78.2%)	2042 (82.9%)	2709 (81.7%)
Diagnosis: Psychological disorders			
Yes	148 (17.4%)	545 (22.1%)	693 (20.9%)
No	663 (77.7%)	1798 (73.0%)	2461 (74.2%)
Diagnosis: Epilepsy			
Yes	10 (1.2%)	25 (1.0%)	35 (1.1%)
No	800 (93.8%)	2315 (94.0%)	3115 (93.9%)
Diagnosis: Kidney disease			
Yes	25 (2.9%)	43 (1.7%)	68 (2.1%)
No	783 (91.8%)	2278 (92.5%)	3061 (92.3%)
Diagnosis: Liver disease			
Yes	21 (2.5%)	42 (1.7%)	63 (1.9%)
No	784 (91.9%)	2290 (93.0%)	3074 (92.7%)
Diagnosis: Degenerative central nervous system disease			
Yes	10 (1.2%)	27 (1.1%)	37 (1.1%)
No	799 (93.7%)	2310 (93.8%)	3109 (93.8%)
Diagnosis: B12 deficiency			
Yes	32 (3.8%)	99 (4.0%)	131 (4.0%)
No	777 (91.1%)	2246 (91.2%)	3023 (91.2%)
Diagnosis: Delirium			
Yes	0 (0%)	4 (.2%)	4 (.1%)
No	800 (93.8%)	2331 (94.6%)	3131 (94.4%)
Diagnosis: Hearing impairment			
Yes	213 (25.0%)	305 (12.4%)	518 (15.6%)
No	602 (70.6%)	2052 (83.3%)	2654 (80.0%)

.35 decrease in PID (CI -.48 to −.22). Primary/secondary education ($\beta = -.01$ [CI -.15 – .12], $R^2 = .03$, $p = .840$) and university education ($\beta = -.03$ [CI -.09 – .04], $R^2 = .03$, $p = .404$) were not significantly associated with PID. In a supplementary model unadjusted for age and sex, primary/secondary education ($\beta = .04$ [CI -.10 – .17], $R^2 = .001$, $p = .586$) and university education ($\beta = -.04$ [CI -.11 – .02], $R^2 = .001$, $p = .174$) were not significantly associated with PID (Supplementary Table 2).

Table 3 – Linear regression models investigating the association of age, sex and education on Propositional Idea Density (PID) in participants of the ISLAND Project. Model 1 adjusted for age. Model 2 adjusted for age and sex. Model 3 adjusted for age, sex and primary/secondary and university education.

Predictors	Propositional Idea Density (PID)								
	Model 1			Model 2			Model 3		
	Estimates	CI	*p*	Estimates	CI	*p*	Estimates	CI	*p*
(Intercept)	53.30	53.17–53.43	**<.001**	53.52	53.36–53.67	**<.001**	53.80	52.21–55.40	**<.001**
Age (in years)	−.40	−.53–−.26	**<.001**	−.35	−.48–−.22	**<.001**	−.41	−.58–−.24	**<.001**
Sex				−.84	−1.14–−.54	**<.001**	−.92	−1.31–−.53	**<.001**
Education (in years)							−.01	−.15 – .12	.840
University Education (in years)							−.03	−.09 – .04	.404
Observations	3316			3316			1925		
R^2/R^2 adjusted	.010/.010			.019/.019			.026/.024		

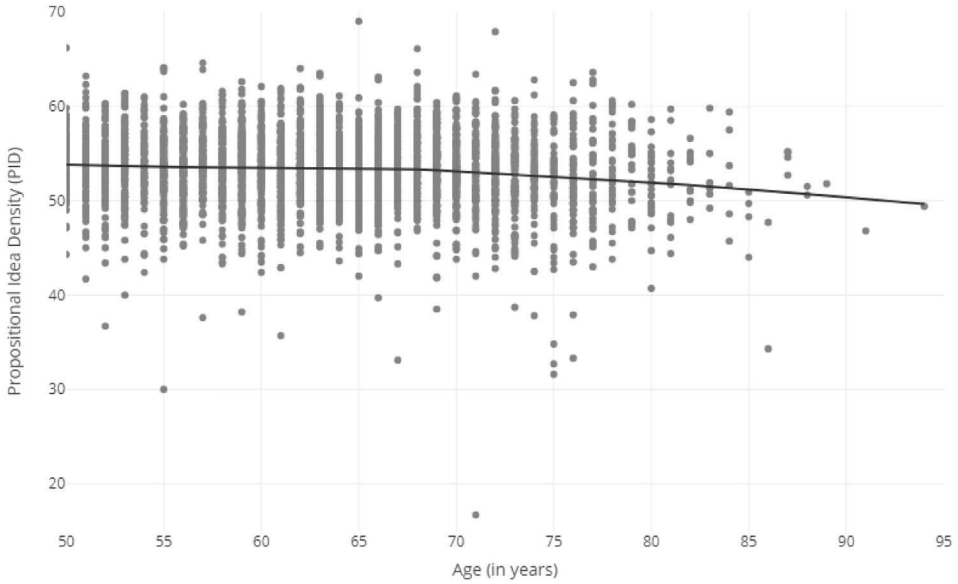

Fig. 1 – Propositional Idea Density (PID) against age in older adult participants of the ISLAND Project (n = 3316). Trend line indicates application of locally estimated scatterplot smoother.

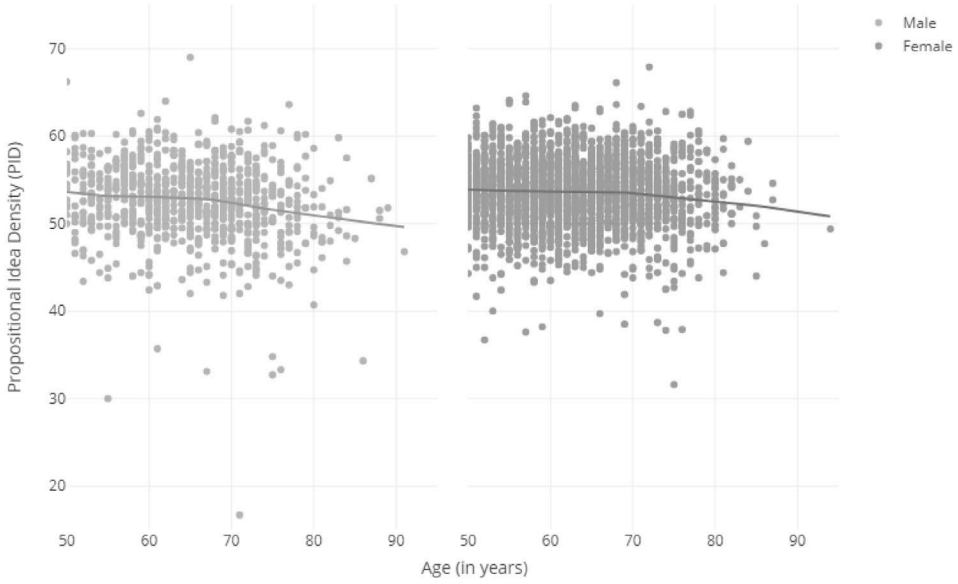

Fig. 2 – Propositional Idea Density (PID) against age in older adult participants of the ISLAND Project, stratified by reported sex (n = 853 male, n = 2463 female). Trend line indicates application of locally estimated scatterplot smoother.

Family history of dementia and all ISLAND participants' medical conditions were not significantly associated with PID (Supplementary Table 3). There was a significant correlation between words and ideas ($r = .99$, $p < .001$), ideas and PID ($r = .13$, $p < .001$) but not words and PID ($r = .01$, $p = .487$). Age (in years) remained significant even when adjusting for words (Supplementary Table 4).

4. Discussion

This cross-sectional study found advancing age and being male were associated with decreased PID collected online via an autobiographical writing task undertaken by participants aged 50 years and over from the ISLAND Project. This confirmed our hypothesis that advancing age, but not education, would be associated with a reduction in PID, in line with several studies (Ferguson et al., 2014; Kemper et al., 2001a, 2001b; Kemper & Sumner, 2001), however our finding with respect to sex was surprising.

Our study found linear decreases in PID with respect to age, in line with Kemper, Greiner, et al. (2001) who reported declines in PID (.30 units per year) at similar rates across a variety of ages. We also observed linear declines in PID measured cross-sectionally in participants over an age range of 50–94 years at approximately .35 PID with each year of advancing age.

Compared with Kemper, Marquis, and Thompson (2001), we saw similar declines with age although our study had a much larger sample size of 3316 participants, with 83 of our participants over the age of 80 years. More recently, Ferguson et al. (2014) investigated longitudinal PID in a large cohort of female participants (n = 19,512) in the Australian Longitudinal Study on Women's Health (Ferguson et al., 2014). For the purpose of analysis, participants were split into three age categories; young adult (18–35 years), mid adult (45–64 years) and older adult (70–87 years). Using CPIDR v3 and linear mixed models, this study reported a significant effect of age for the older participant group. Ferguson et al. (2014) reported a drop in PID of the order of .2 per year of follow up, compared with 2.0 per year reported in Kemper, Marquis, and Thompson (2001). Our study reported an average reduction of .36 PID with each year of advancing age, although this significant association with age was cross-sectional in nature. It remains to be seen whether our results hold up with longitudinal measures of PID.

Our findings were similar to previous research in mostly female cohorts (Ferguson et al., 2014; Kemper, Greiner, et al., 2001). Given our primarily female cohort (74.3%), this may explain the similarity of findings with respect to declines in PID, although we also observed sex as a significant predictor of PID. To our knowledge, this is the first study to investigate PID in a large-scale cohort of males and females (>200 individuals). Given sex disparities in AD (Ferretti et al., 2018) and a general effect of sex on language function (Ullman et al., 2008), we may be witnessing early signs of divergence in PID between the sexes. Our cross-sectional study was not able to investigate whether these disparities remain over time, however further research is planned to investigate whether the observed differences with respect to sex result in divergent trajectories over time.

In accordance with our hypothesis, our study did not find an association between education (primary/secondary school and university) and PID. Snowdon et al. (1996) investigated written discourse from 93 participants in the Nun Study with similar average education (Nun Study: 17 years, ISLAND: 16 years), and found a significant association between PID and years of education ($r = .27$, $p < .01$). Given sisters displayed considerable homogeneity in social, occupational, nutritional and medical experiences over the years, the Nun study can control for many confounding lifestyle factors. Therefore, our null findings with respect to education could be due to an unknown contributor of variance. Interestingly, the Nun study found neuropathological evidence of AD in all participants with low PID and in no participants with high PID. Biological markers of AD may therefore be associated with linguistic aspects of cognition, however more research is needed to quantify this relationship. In line with our findings, several studies have reported non-significant results between education and PID (Farias et al., 2012; Kemper & Sumner, 2001). Although Kemper and Sumner (2001) reported significant PID differences with respect to age groups, no significant difference was found for education. Farias et al. (2012) reported a modest result ($r = .17$, $p = .12$) between PID and education in 81 participants (mean education 14.4 [±3.6] years). Given the majority of studies have reported a lack of association with education, PID may be measuring different aspects of cognitive reserve that fluctuate irrespective of educational engagement. Further longitudinal research into the cognitive and biomarker hallmarks of AD with PID would reveal the direction of the effect and illuminate whether PID may represent a candidate marker for online screening tools at the population level.

We acknowledge our study is not without limitations. Current participants in the ISLAND Project represent a highly educated cohort, with both primary/secondary and university education above state population norms (Australian Bureau of Statistics, 2016). We acknowledge how this bias may impact our results. We did not observe an association between family history of dementia or other medical comorbidities and PID, however this may be in part due to healthy cohort bias inherent in our sample. Longitudinal follow-up over the course of the project may reveal associations between PID and both modifiable and non-modifiable risk factors for AD. ISLAND recruitment is ongoing, and we hope to recruit a more heterogenous cohort with respect to educational, occupational and sex stratification. This was a cross-sectional investigation, with a single time point of autobiographical writing text collection. Our effect sizes were relatively small, and we acknowledge the possibility our sample was overpowered which could impact replicability. Further, we were also limited by barriers to participation such as computer literacy, internet access and self-selection bias. Probes given to ISLAND participants regarding the content of their autobiographies were also unable to be controlled for yet were presented identically to all ISLAND participants. Longitudinal analysis is planned, with biennial autobiographical collection through each participant's online ISLAND Home Portal in preparation. We also plan to investigate associations between PID and candidate blood-based biomarkers of AD and dementia.

5. Conclusion

In this cross-sectional study of a longitudinal cohort, we found increasing age and being male were significantly associated with a decrease in PID. Education, both primary/secondary and university, was not associated with PID. Although our effect sizes were small, this study contributes to a growing field of research investigating language as a potential early detection tool for AD and other forms of dementia. We observed significant associations of PID with non-modifiable risk factors for AD (age and sex), yet none of the modifiable risk factors included in our study (education, family history). PID may therefore represent an ideal measure of cognitive reserve that fluctuates independent of education. Administered via the ISLAND Home Portal, this study demonstrates that an autobiographical writing task is a valid method of online data collection in older adults, and repeat measurement is planned. This study indicates collecting autobiographical texts and utilizing automated calculation of PID (using CPIDR v5.1) is a potential marker of cognition measured through language that can be easily collected at the population level through online administration.

The ISLAND Project is a longitudinal investigation into the amelioration of dementia risk through modifiable risk factors, and will provide valuable insight to how positive behavioural and lifestyle change can protect against the cognitive and biomarker hallmarks of dementia. Over the 10 years of the project, oral and written language measures will be collected every two years, enabling a longitudinal investigation into

individual change in PID, and how partaking in planned interventions targeting dementia risk reduction may relate to these linguistic indicators of cognitive function.

Author contributions

Edward Hill: Conceptualization, Methodology, Formal Analysis, Visualization, Writing—Original Draft.

Jane Alty: Conceptualization, Writing—Review & Editing.

Larissa Bartlett: Writing—Review & Editing, Project administration.

Lyn Goldberg: Writing—Review & Editing.

Mira Park: Writing—Review & Editing.

Soonja Yeom: Writing—Review & Editing.

James Vickers: Supervision, Conceptualization, Funding acquisition, Writing—Review & Editing.

Data availability statement

The conditions of our ethics approval prevent public archiving of the ISLAND data and code supporting the current study. Readers seeking access to this data should direct requests to the corresponding author, e.hill@utas.edu.au. Data will be made available subject to completion of a data transfer agreement.

Declaration of competing interest

The authors declare no financial or other conflicts of interests.

Acknowledgements

The authors are grateful for contributions made by the ISLAND Project team: Helen Douglas, Dr Adam Kane and Justine Keay; and technical and data management by Timothy Saunder, Chris Parker, Joshua Eastgate, Oliver Freeman and Jay Borchard. We are indebted to the participants of the ISLAND Project, who make this work possible. This work was supported by the Australian Government's Medical Research Futures Fund—Keeping Tasmanians out of Hospital, Masonic Medical Research Foundation Limited, St Luke's, Terry and Maureen Hopkins Foundation and the JO & JR Wicking Trust (Equity Trustees). We gratefully acknowledge the benefits to this paper of reviewers' comments as part of the peer review process.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2021.09.018.

REFERENCES

Australian Bureau of Statistics. (2016). *Census data. 2016 census QuickStats*. Australia: Australian Bureau of Statistics. https://www.abs.gov.au/websitedbs/censushome.nsf/home/2016. (Accessed 1 March 2021).

Bartlett, L., Doherty, K., Farrow, M., Kim, S., Hill, E., King, A., Alty, J., Eccleston, C., Kitsos, A., Bindoff, A., & Vickers, J. (2021). The ISLAND study linking ageing and neurodegenerative disease: Protocol and progress. *Journal of Medical Internet Research - Research Protocols*. Submitted for publication.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods, 40*(2), 540–545.

Christensen, H., Anstey, K., Leach, L., & Mackinnon, A. (2008). Intelligence, education, and the brain reserve hypothesis. In *The handbook of aging and cognition* (3rd ed., pp. 133–189). United Kingdom: Psychology Press, Taylor & Francis.

Clarke, N., Foltz, P., & Garrard, P. (2020). How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer's disease. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 129*, 449–463. https://doi.org/10.1016/j.cortex.2020.05.001

Engelman, M., Agree, E. M., Meoni, L. A., & Klag, M. J. (2010). Propositional density and cognitive function in later life: Findings from the Precursors Study. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 65*(6), 706–711. https://doi.org/10.1093/geronb/gbq064

Farias, S. T., Chand, V., Bonnici, L., Baynes, K., Harvey, D., Mungas, D., Simon, C., & Reed, B. (2012). Idea density measured in late life predicts subsequent cognitive trajectories: Implications for the measurement of cognitive reserve. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 67*(6), 677–686. https://doi.org/10.1093/geronb/gbr162

Ferguson, A., Spencer, E., Craig, H., & Colyvas, K. (2014). Propositional Idea Density in women's written language over the lifespan: Computerized analysis. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 55*, 107–121. https://doi.org/10.1016/j.cortex.2013.05.012

Ferretti, M. T., Iulita, M. F., Cavedo, E., Chiesa, P. A., Dimech, A. S., Chadha, A. S., Baracchi, F., Girouard, H., Misoch, S., Giacobini, E., Depypere, H., Hampel, H., & Women's Brain Project & Alzheimer Precision Medicine Initiative. (2018). Sex differences in Alzheimer disease—the gateway to precision medicine. *Nature Reviews Neurology, 14*(8), 457–469. https://doi.org/10.1038/s41582-018-0032-9

Hebert, L. E., Weuve, J., Scherr, P. A., & Evans, D. A. (2013). Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology, 80*(19), 1778–1783. https://doi.org/10.1212/WNL.0b013e31828726f5

Kemper, S., Greiner, L. H., Marquis, J. G., Prenovost, K., & Mitzner, T. L. (2001). Language decline across the life span: Findings from the nun study. *Psychology and Aging, 16*(2), 227–239. https://doi.org/10.1037/0882-7974.16.2.227

Kemper, S., LaBarge, E., Ferraro, F. R., Cheung, H., Cheung, H., & Storandt, M. (1993). On the preservation of syntax in Alzheimer's disease: Evidence from written sentences. *Archives of Neurology, 50*(1), 81–86. https://doi.org/10.1001/archneur.1993.00540010075021

Kemper, S., Marquis, L., & Thompson, M. (2001). Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging, 16*(4), 600–614. https://doi.org/10.1037//0882-7974.16.4.600

Kemper, S., & Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and Aging, 16*(2), 312–322.

Kim, B. S., Kim, Y. B., & Kim, H. (2019). Discourse measures to differentiate between mild cognitive impairment and healthy aging. *Frontiers in Aging Neuroscience, 11*, 221. https://doi.org/10.3389/fnagi.2019.00221

Komeili, M., Pou-Prom, C., Liaqat, D., Fraser, K. C., Yancheva, M., & Rudzicz, F. (2019). Talk2Me: Automated linguistic data collection for personal assessment. *Plos One, 14*(3), Article e0212342. https://doi.org/10.1371/journal.pone.0212342

Laguarta, J., Hueto, F., Rajasekaran, P., Sarma, S., & Subirana, B. (2020). Longitudinal speech biomarkers for automated Alzheimer's detection. *Frontiers in Computer Science, 18*(3), Article e624694. https://doi.org/10.3389/fcomp.2021.624694

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S., Fox, N., Gitlin, L., Howard, R., Kales, H., Kivimäki, M., Larson, E., Ogunniyi, M., … Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the lancet commission. *Lancet, 396*, 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6

Lyons, K., Kemper, S., LaBarge, E., Ferraro, F. R., Balota, D., & Storandt, M. (1994). Oral language and alzheimer's disease: A reduction in syntactic complexity. *Aging, Neuropsychology, and Cognition, 1*(4), 271–281. https://doi.org/10.1080/13825589408256581

Morris, R. G. (1996). *The cognitive neuropsychology of Alzheimer-type dementia*. United Kingdom: Oxford University Press.

Riley, K. P., Snowdon, D. A., Desrosiers, M. F., & Markesbery, W. R. (2005). Early life linguistics ability, late life cognitive function, and neuropathology: Findings from the nun study. *Neurobiology of Aging, 26*(3), 341–347. https://doi.org/10.1016/j.neurobiolaging.2004.06.019

Snowdon, D. A., Kemper, S., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Jsls: Journal of the Society of Laparoendoscopic Surgeons, 275*(7), 528–532.

Turner, A., & Greene, E. (1997). *The construction and use of a propositional text base*. United States of America: Institute for the Study of Intellectual Behavior, University of Colorado Boulder.

Ullman, M. T., Miranda, R. A., & Travers, M. L. (2008). Sex differences in the neurocognition of language. In J. B. Becker, K. J. Berkley, E. Hampson, J. P. Herman, & E. A. Young (Eds.), *Sex differences in the brain: From genes to behavior* (pp. 291–309). Cary, NC: Oxford University.