when all predictors were entered into the same model, visuospatial performance did not significantly contribute to the prediction of pillbox errors. These results suggest that providers may benefit from investigating medication management abilities when deficits in PS, WM, and/or language are identified. Further research is needed to better understand which domains best predict PT failure.

Categories:
Assessment/Psychometrics/Methods (Adult)
Keyword 1: activities of daily living
Correspondence: Frances M. Bozsik, Memphis VA Medical Center, frances.bozsik@gmail.com

---

17 The Chinese Version of Craft Story Recall: A Preliminary Study on the Diagnostic Values of Mild Cognitive Impairment and Dementia.

Gelan Ying[1], Judith Neugroschl[2], Amy Aloysi[2], Dongming Cai[2,3], Tianxu Xia[2], Carolyn W Zhu[2,3], Xiaoyi Zeng[2], Jimmy Akrivos[2], Linghsi Liu[2], Yiyu Cao[2], Wei-Qian Wang[2], Mary Sano[2,4], Glenn E Smith[1], Clara Li[2]
[1]University of Florida, Gainesville, FL, USA.
[2]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]James J. Peters VA Medical Center, Bronx, Bronx, NY, USA. [4]James J. Peters Veterans Affairs Medical Center, Bronx, NY, USA

Objective: Craft story recall test in the National Alzheimer's Coordinating Center Uniformed Data Set 3 (NACC UDS3) neuropsychological battery has been employed to assess verbal memory and assist clinical diagnosis of mild cognitive impairment (MCI) and dementia. While a Chinese version of the test was adapted, no existing literature has examined the diagnostic validity of the test in Chinese Americans. This study aimed to evaluate the predictive validity of both immediate and delayed recall.
Participants and Methods: The Chinese version of Craft Story was administered in to 78 Chinese participants per their language preference of Mandarin or Cantonese. Outcome measures were verbatim and paraphrase recall of the story immediately and after a 20-minute delay. A multiple linear regression was performed to investigate the association of each outcome measure with age, education, gender, age when moved to the U.S., years in the U.S., and testing language. To assess its diagnostic value, cutoff standard deviation scores of -1.5 and -2.0 from the mean of the clinically cognitive normal participants were generated for MCI and dementia diagnoses, respectively. Due to the small sample size, a normative group fitting the mean age (73 years), years of education (12 years), and the majority gender (female) of the current sample were used to identify standard cut points. A receiver-operating characteristic analysis was used to compare predicted diagnosis with actual clinical diagnosis obtained through patients' overall performance and a consensus meeting by licensed clinicians.
Results: Younger age ($p < 0.05$) and being tested in Mandarin ($p < .01$) were positively associated with immediate and delayed recall. Strong positive correlations between each measure were observed (all $p < .001$), indicating a significant relationship between information encoded and retained. Among all the participants, 15 (19.2%) were diagnosed with MCI and 22 (28.2%) with dementia. For MCI diagnosis, the standard cutoff scores demonstrated adequate sensitivity (verbatim=82%, paraphrase=91%) but low specificity (verbatim=44%, paraphrase=67%) in all outcome measures. For dementia diagnosis, delayed recall showed strong sensitivity (100%) and adequate specificity (75%) in both verbatim and paraphrasing scores. Immediate recall paraphrase (sensitivity = 95%, specificity = 50%) showed a better sensitivity but lower specificity than verbatim scoring (sensitivity = 86%, specificity = 58%). The accuracy was higher in delayed recall for both MCI and dementia diagnosis. A preliminary analysis on the optimal cut points indicated higher cutoff scores to distinguish MCI and dementia from clinically cognitive normal population, and from each other (e.g., the optimal cut point for delayed verbatim in distinguishing MCI from normal is 8.0 (sensitivity=89%, specificity=73%, AUC=84.3%)).
Conclusions: Consistent with previous literature, Craft Story delayed recall served as a more accurate diagnostic tool for both MCI and dementia compared to immediate recall in older Chinese Americans. However, poor specificity might increase the chance of following false positive subjects in clinical trials. In addition, testing language appeared to impact performance on verbal memory recall of constructed information. Thus, future studies

should focus on developing normative scores that address both the overall cultural differences of Chinese Americans and the heterogeneity within this population.

---

18 Automatically calculated lexical and sentential context features of connected speech predict cognitive impairment

Graham Flick[1,2], Rachel Ostrand[2]
[1]New York University, New York, New York, USA. [2]IBM Research, Yorktown Heights, New York, USA

Objective: Early detection is critical to the effective management of Alzheimer's Disease (AD) and other forms of dementia. Existing screening assessments are often costly, require substantial expertise to administer, and may be insensitive to mild changes in cognition. A promising alternative is automatically measuring features of connected speech (c.f., Ostrand & Gunstad, 2021, Journal Geriatric Psych & Neurol) to predict impairment. Here, we built on prior work examining how well speech features predicted cognitive impairment. Unique to the current work, we attempted to capture more holistic effects of cognitive impairment by examining the relevance of linguistic features that measure sentential or discourse context properties of speech, including the context in which filler words (e.g., um) occur, and the predictability of individual words within their sentence context, computed from a large computational language model (GPT-2).
Participants and Methods: Participants completed the Cookie Theft picture description task, with data available in the DementiaBank corpus (Becker et al., 1994, Arch Neurol). Descriptions that contained at least 50 words (N = 214) were submitted to an automatic feature calculation pipeline written in Python to calculate various part-of-speech counts, lexical diversity metrics, and mean lexical frequency, as well as multiple metrics related to lexical surprisal (i.e., how surprising a word is given its context).

Surprisal of individual words was computed using the pre-trained GPT-2 transformer language model (Radford et al., 2019, Comput. Sci.) by computing word probability given the previous 12 words. Multiple linear regression was performed using 17 linguistic features jointly as predictors, and Mini-Mental State Examination (MMSE) score as the outcome variable. Simple regressions were calculated between each feature and MMSE scores to examine the predictability of each linguistic feature on cognitive decline.
Results: A multiple linear regression model containing all linguistic features plus demographic information (age, sex, education) significantly predicted MMSE scores (Adjusted $R^2 = 0.41$, $F_{20, 193} = 8.37$, $p < .001$), and explained significantly more variance in MMSE scores than did demographic variables alone ($F_{17,193} = 6.85$, $p < .001$). Individual predictors that were significantly correlated with MMSE score included: how unexpected an individual's word choices were, given the preceding context (median surprisal: $r = -0.33$, $p < 0.001$; interquartile range: $r = 0.18$, $p < 0.02$), mean lexical frequency ($r = -0.50$, $p < .001$), and usage of definite articles ($r = 0.31$, $p < 0.001$), nouns ($r = 0.26$, $p < .001$), and empty words ($r = -0.25$, $p < 0.001$).
Conclusions: Participants with lower MMSE scores, indicating greater impairment, used more frequent, yet more surprising, words, and produced more empty words and fewer definite articles and nouns. These results suggest that measures of semantic specificity and coherence of speech could be meaningful predictors of cognitive decline, and can be computed automatically from speech transcriptions. The results also provide novel evidence that computational approaches to estimating lexical predictability may have value in predicting the degree of decline, motivating future work in other speech elicitation tasks and differing clinical groups.