Random Item Generation Is Affected by Age

Namita Multani,a Frank Rudzicz,b,c,d Wing Yiu Stephanie Wong,a Aravind Kumar Namasiyayam,a,d and Pascal van Lieshouta,b,c,d,e

Purpose: Random item generation (RIG) involves central executive functioning. Measuring aspects of random sequences can therefore provide a simple method to complement other tools for cognitive assessment. We examine the extent to which RIG relates to specific measures of cognitive function, and whether those measures can be estimated using RIG only.

Method: Twelve healthy older adults (age: M = 70.3 years, SD = 4.9; 8 women and 4 men) and 20 healthy young adults (age: M = 24 years, SD = 4.0; 12 women and 8 men) participated in this pilot study. Each completed a RIG task, along with the color Stroop test, the Repeatable Battery for the Assessment of Neuropsychological Status, and the Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007). Several statistical features extracted from RIG sequences, including recurrence quantification, were found to be related to the other measures through correlation, regression, and a neural-network model.

Results: The authors found significant effects of age in RIG and demonstrate that nonlinear machine learning can use measures of RIG to accurately predict outcomes from other tools.

Conclusions: These results suggest that RIG can be used as a relatively simple predictor for other tools and in particular seems promising as a potential screening tool for selective attention in healthy aging.

The ability to generate sequences randomly has been used to predict important aspects of central executive functioning, because it requires active generation of new strategies and inhibition of stereotypical responses (Baddeley, 1966). For random item generation (RIG), a participant is typically asked to produce a random sequence of items such as letters of the alphabet, numbers, or keystrokes. In general, humans demonstrate greater difficulty in producing random series of items, compared with computer-generated random series (Ginsburg & Karpiuk, 1994; Rabinowitz, 1970; Spatt & Goldenberg, 1993). When associated with highly overlearned associations (e.g., ascending sequences of numbers or letters), this capacity for randomness may depend on two main factors: (a) the ability to inhibit competing or habitual responses (i.e., suppressing stereotyped sequences) and (b) the ability to allocate attention and working-memory resources to monitoring and updating responses according to a perceived concept of randomness (Miyake et al., 2000). Both of these factors place demands on executive function, which may relate to prefrontal cortical functions (Baddeley, Emslie, Kolodny, & Duncan, 1998; Miyake et al., 2000). Thus, producing random sequences of items relates to central executive function, because generating long random sequences is associated with neither short-term memory deficits nor misunderstood notions of randomness (Baddeley, 1998; Wagenaar, 1970).

RIG tasks have been applied in various clinical populations, including people with autism, Asperger’s disorder, brain injury, and Alzheimer’s disease (Breidt, 1973; Brugger, Monsch, Salmon, & Butters, 1996; Rinehart, Bradshaw, Moss, Berettoni, & Tonge, 2006). Compared with healthy adults, people with neuropsychological deficits display decreased performance in generating random patterns of numbers during the execution of such tasks. In a study by Brugger et al. (1996), individuals with Alzheimer’s disease were compared with a healthy control group on their ability to generate random numbers. That analysis showed that people with Alzheimer’s disease were more likely to generate items outside of the allowable set or vocabulary, and with a higher number of consecutive digit pairs (Evans, 1978), which indicates a more stereotypical response pattern (Brugger et al., 1996). This evidence relates...
the pronounced declines in central executive processes with the reduced ability to produce sequences of random items in Alzheimer’s disease (Morris, 1994).

Peters, Giesbrecht, Jelicic, and Merckelbach (2007) independently discovered a positive relationship between Stroop interference (used to measure selective attention) and seriation, which is a process defined by generating stereotypical schemas (Williams, Moss, Bradshaw, & Rinehart, 2002). In other words, problems of selective attention are negatively correlated with the ability to produce random sequences. Jahanshahi, Saleem, Ho, Dinrberger, and Fuller (2006) also suggested that the ability to inhibit learned schemas as a reflection of selective attention capacity is the fundamental component underlying RIG performance.

Studies of RIG performance are not limited to patient populations. Van der Linden, Beerten, and Pesenti (1998) performed a RIG study on healthy older adults and found that they produced less-random sequences compared with their younger counterparts and that this difference grew with increasing speaking rate. Furthermore, compared with young adults, older adults displayed greater decline in randomness with increasing cognitive load, including under dual-tasking conditions. This suggests that age modulates the ability to allocate cognitive resources, which is consistent with the literature (e.g., Glisky, 2007).

A few additional studies have examined random number generation in healthy young and older adults. These studies also have demonstrated an age-related decline in the ability to generate random series of numbers, related to a reduced ability in older adults to suppress habitual information (counting, subtraction and addition), and this may be exaggerated by a decline in processing speed (Albinet, Boucard, Bouquet, & Audiffren, 2012; Albinet, Tomporowski, & Beasman, 2006; Boucard et al., 2012; Maylor & Wing, 1996).

These studies indicate the potential of using RIG as a tool to identify issues of central executive functioning in healthy older adults, in particular with respect to selective attention and inhibition processes (Van der Linden et al., 1998). The color Stroop task (Stroop, 1935) and specific subtests within the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph, 2012) have traditionally been used to study central executive processing in healthy older adults (Randolph, Tierney, Mohr, & Chase, 1998). In addition, the Peabody Picture Vocabulary Test–Fourth Edition (PPVT-4; Dunn & Dunn, 2007) has been used to measure verbal intelligence in older adults (Bialystok & Luk, 2012). However, these tasks can be time consuming (e.g., the RBANS takes 30 min or more) and require trained professionals to administer and interpret their results. More important, these tests are sensitive to lexical access, lexico-semantic impairments, and sensory processing issues in older adults and patient populations. For example, in individuals with Alzheimer’s disease a significant portion of the color Stroop effects ($R^2 = .87$) can be attributed to sensory degradation (i.e., degraded color vision) and not a decline in selective attention (Ben-David, Tewari, Shakuf, & van Lieshout, 2014), thus highlighting the need to control for such variables in testing cognitive performance in older populations.

Given that RIG can be completed in 5 min and is less sensitive to sensory acuity and lexico-semantic processing issues, we explored the potential of RIG as a tool to identify issues in particular aspects of central executive functioning in older adults, avoiding confounds related to sensory degradation and/or lexical-processing issues. To be specific, we evaluated how RIG performance relates to outcomes on well-established measures of central executive functioning, namely the PPVT-4, RBANS, and color Stroop task. A secondary purpose of this study was to explore the benefit of using nonlinear analyses as an alternative to more traditional linear statistical techniques such as correlation, because the relationship between outcome measures on different tasks may not be linear. We used a state-of-the-art (LeCun, Bengio, & Hinton, 2015) “deep” artificial neural network that predicts submeasures in each framework (e.g., Stroop) given relevant statistics derived from RIG sequences.

Method

Data Collection

Participants

Twelve healthy older adults (age: $M = 70.3$ years, $SD = 4.9$; eight women and four men) and 20 healthy young adults (age: $M = 24$ years, $SD = 4.0$; 12 women and eight men) participated in this experiment. All participants reported good health, spoke English as a first language, and completed the PPVT-4 and the RBANS. The PPVT-4 measures verbal intelligence by instructing participants to point to an appropriate image out of four alternatives for each auditory stimulus word. This task has standardized scores, and available normative data extend up to age 70 years. The RBANS measures immediate and delayed memory function, attention, and visuospatial and language functions. In addition, participants performed a variant of the color Stroop task and RIG, both described in the following sections.

Color Stroop Task

The color Stroop task consists of four parts, with separate blocks for each task and 12 trials per block (Stroop, 1935). Participants were seated in front of a 17-in. flat color monitor. The order of stimuli in each block was randomized, with a different random order for each block and for each participant. The word stimulus consisted of four color names (red, blue, yellow, and green) and a string of letters (oooo). The font was 28-point Times New Roman, center aligned, for all computer tasks. For the reading block, participants saw color names that appeared in a white font, and they were asked to read the words aloud as quickly and accurately as possible. In the other blocks (incongruent, baseline, congruent), participants were instructed to name the color in which the word appeared.
as quickly and accurately as possible. For the incongruent block, participants were presented with color names that appeared in nonmatching colors. In the baseline block, the string \( \text{o000o} \) was presented in different colors across stimuli. In the congruent block, the color names appeared in their corresponding colors. Before each block, participants received instructions and two practice trials to familiarize themselves with the task. All stimuli were displayed on a black background. Each stimulus presentation was triggered by the participant’s response as picked up by a Blue Snowball microphone. Individual scores which were self-scored were excluded from the data analysis. Mean reaction time was used as a measure of performance (individual scores that were inaccurate or self-corrected were excluded from mean reaction time). The interference was calculated as the difference between congruent and incongruent trials.

**RIG**

The RIG task was recorded using a Zoom H4 hand-held audio recorder. At the beginning of the RIG task, participants were given the following instruction:

Please imagine that you are drawing letters from a hat, one at a time, and calling them out loud and putting them back into the hat. We want a random non-repeating set of English sounds. Please be reminded that by using this method it is very unlikely that letter strings would be in alphabetic sequences and should not result in English words. There should also not be an excess of the same letter or combination of letters in the same order. You have to keep repeating this for between 3 and 5 minutes.

We imposed no fixed constraint to the rate of RIG, to account for possible differences in processing speed across age groups (Ben-David et al., 2014). We derived several measures of participant-generated sequences to see (a) what aspects of their performance were most affected by age and (b) what aspects were indicative of changes in specific functions as measured in other tests. For example, the RIG task provides a means to evaluate excessive repetitiveness, which can indicate a lack of selective attention (Jahanshahi et al., 2006). To explore these functions more specifically, several novel features are derived from a recurrence quantification analysis (RQA) of those sequences. These features quantify statistical aspects of recurrent patterns over time (Little, McSharry, Roberts, Costello, & Moroz, 2007; Marwan, Romano, Thiel, & Kurths, 2007) and can therefore capture the extent to which a speaker resorts to repetitive patterns of characters. From RQA we derive the following features:¹

- **Recurrence rate:** the density of recurrence points in a density plot
- **Determinism:** the percentage of recurrence points forming diagonal lines (a measure of predictability)
- **Mean diagonal-line length:** a measure of the extent of predictable repetitions
- **Maximal diagonal-line length (LMAX):** the longest sequence
- **Entropy of diagonal-line lengths:** a measure of the complexity of the determinist structure in the system (computed over the distribution of diagonal-line lengths)
- **Laminarity:** a measure of intermittency related to laminar phases (computed over vertical lines in the recurrence plot)
- **Trapping time:** the average length of vertical lines in the recurrence plot
- **Maximal vertical-line length:** the maximum of these measures
- **Recurrence time of first type:** times between subsequent points in the recurrence plot (see Gao & Cai, 2000)
- **Recurrence time of second type:** same as the first type, but with subsequent points within the sphere of analysis removed; as with the first type, this is related to the information content of the attractor
- **Recurrence-period density entropy (RPDE):** a measure of repetitiveness in a signal, mimicking entropy, on a \([0, 1]\) scale
- **Clustering coefficients:** the degree to which points in the recurrence plot cluster together
- **Transitivity:** the total number of points connecting any two sequences

Several other features are derived from probability distributions obtained from histograms over all characters (unigraphs) or pairs of adjacent characters (bigraphs) in the response patterns. Flatter distributions imply that all characters (or pairs of characters) are being selected with increasing consistency, which implies true uniform randomness and a good degree of selective attention. Statistical entropy is a measure of the flatness of a distribution; therefore, it is computed for each sequence, for both unigraphs and bigraphs. Other aspects of these distributions characterize their shape. To be specific, we measured kurtosis to determine whether excessive peaks in these distributions imply an inability to distribute choices evenly; we also measured skewness, to quantify the asymmetry of the distributions, which would imply clustering effects. Moreover, because we were interested in examining individual deviation from group performance, individual distributions were compared against the respective unigraph and bigraph distributions generated from all other speakers using standard analysis techniques, namely Kullback–Leibler divergence, Cohen’s \(d\) effect size, and the \(F\) test. These measures are specified in the online supplemental materials.

¹Details are in the online supplemental materials. Source code is available at http://tocsy.pik-potsdam.de/CRPtoolbox/.
Relating RIG With Other Tools

We examined the extent to which RIG-based statistics can be used to predict submeasures in other tools that are typically used to measure aspects of cognitive function. For this, we applied a mixture-density neural-network (MDN) method, which is based on state-of-the-art discriminative machine learning. To provide context, we also applied multilinear regression as a baseline. The MDN is a feed-forward neural network whose output provides the parameters for a mixture of Gaussians (Richmond, King, & Taylor, 2003). The number of RIG features \(|x|\) and number of output Gaussians \(M\) are independent empirical parameters, and the number of hidden units is proportional to each—that is, \((1/2)(x + 3M)\). The mixture model is linear in that there is a single covariance parameter for each component. The MDN is initialized with 10 iterations of \(k\)-means clustering and subsequently trained up to a maximum of 200 iterations with scaled conjugate gradient optimization.

In all cases, we used leave-one-out cross-validation in which we predicted submeasures of standardized assessments for each individual, in turn, given models tuned using data from all other individuals in the data set. We use the standard relative error formulation to evaluate accuracy of predictions, as described in the online supplemental materials.

Results

Scores on the RBANS are summarized in Table 1. On average, standard Stroop scores did not differ significantly, \(t(30) = 0.63, p = .53\), between older adults \((M = 111.8, SD = 14.3)\) and younger adults \((M = 109.0, SD = 10.9)\). Table 2 shows pairs of variables correlated between RIG and the other measures. From this perspective, RIG is highly correlated with Stroop, as visualized in Figure 1, and to some extent with the subscores of the RBANS (especially List Recognition and Semantic Fluency), but not with the PPVT-4. To be specific, both the Stroop Interference (difference between reaction times in congruent and incongruent stimuli) and the total Stroop effect are highly negatively correlated with both bigraph entropy and transitivity; that is, people who have more difficulty with incongruent stimuli in the Stroop task are also much more likely to repeat themselves in RIG.

In order to measure the possible effects of age on assessments, two-sample, two-tailed homoscedastic \(t\) tests were performed across the younger and older groups for each feature in each tool under consideration (see the online supplemental materials). Among measurements in the Stroop task, significant effects of age with large effect sizes (Cohen's \(d\)) were noted for both interference, \(t(29) = 4.1, p < .001, 95\% CI [178, 412], d = −1.64\), and the total effect, \(t(29) = 4.6, p < .001, 95\% CI [199, 512], d = −1.47\). Among RBANS subscores, significant effects of age with small to medium effect sizes were noted for List Learning (part of Immediate Memory), \(t(29) = −3.7, p < .001, 95\% CI [−8.4, −2.5]\), \(d = 0.14\), and Coding (part of Attention), \(t(29) = −5.9, p < .001, 95\% CI [−24, −12], d = −0.46\). No other features from the PPVT-4 \((d = −0.12)\), Stroop task (Facilitation subscore \(d = 0.52\)), RBANS (total scale \(d = −0.31\)), or RBANS subscores (Visuospatial \(d = −0.84\); Language \(d = 0.74\); Delayed Memory \(d = −0.26\)) showed significant differences across age groups at \(α = .001\), with Bonferroni correction in all cases.

<table>
<thead>
<tr>
<th>Score</th>
<th>Young adults</th>
<th>Older adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immediate Memory</td>
<td>101.35</td>
<td>99.67</td>
</tr>
<tr>
<td>Total</td>
<td>99.55</td>
<td>103.73</td>
</tr>
<tr>
<td>Visuospatial</td>
<td>88.80</td>
<td>104.00</td>
</tr>
<tr>
<td>Language</td>
<td>102.10</td>
<td>93.75</td>
</tr>
<tr>
<td>Attention</td>
<td>110.80</td>
<td>118.67</td>
</tr>
<tr>
<td>Delayed Memory</td>
<td>96.40</td>
<td>99.18</td>
</tr>
<tr>
<td>Total</td>
<td>99.55</td>
<td>103.73</td>
</tr>
</tbody>
</table>

Discussion

We first demonstrated that measures derived from RIG can be highly dependent on age. However, the data also showed several strong linear relationships between RIG and standard tests of cognitive function, and those results can, to a significant degree, be accurately predicted by modern machine-learning methods.

To our knowledge, this study is the first to demonstrate relationships between performance on a verbal (non-lexical) version of the RIG task and standard cognitive assessments in a healthy older population. To be specific, we showed a positive relationship between RIG and specific subcomponents of the RBANS related to semantic fluency and list recognition, and between RIG and the total Stroop effect. We also found significant effects of age on RIG and, to a lesser extent, for Stroop Interference and total effect and the RBANS List Learning and Coding scores. Overall, this demonstrates that analysis of recurrence patterns (i.e., measures of RQA, outlined earlier and detailed in the online supplemental materials) and statistical distributions in RIG can be used to predict performance.
on other tools related to cognitive executive functions with high accuracy using state-of-the-art machine learning. Our intention hereafter is to validate our implementation by exploring these associations with larger populations and focusing on selective attention. Given the changes that occur in RIG due to age-related processes (Van der Linden et al., 1998), our results emphasize the need to account for age effects in studies involving healthy older adults, including studies involving populations with age-related cognitive disorders such as dementia.

We have shown that RIG relates to selective attention processes in that it correlates strongly with Stroop...
results, a tool that has been validated for assessing those processes (Peters et al., 2007). To some extent, this suggests that RIG can be used as a simple screening tool in clinical settings to assess this aspect of cognitive function in particular.

The effect of age on selective attention processes can be interpreted in the context of the inhibitory deficit hypothesis (Hasher & Zacks, 1988). According to this hypothesis, the ability of healthy older adults to ignore irrelevant information is impaired by age-related declines in inhibitory processes. Here, task-irrelevant information enters working memory and hinders processing of task-relevant information. Therefore, age-related differences in RIG may be a result of a limited efficiency of inhibitory processes. In contrast, we found no association between RIG and the PPVT-4, suggesting that RIG is not affected by differences in vocabulary or comprehension of lexical semantics.

Healthy older adults show greater adjacency scores (which depict the distribution of neighboring digits) and longer reaction times on random number generation and the Stroop task, respectively (Boucard et al., 2012), than healthy younger adults. Our preliminary findings support the notion that RIG can serve as a measure for related inhibitory processes.

It is important to note that although inhibition plays a significant role during RIG performance, it is not the only process involved. The ability to successfully generate random items is also contingent upon other factors, including monitoring to ensure random generation (Sexton & Cooper, 2014). Furthermore, the executive subprocess model proposed by Sexton and Cooper proposes that inhibition during RIG is a complex phenomenon that requires interplay among several processes within a limited time frame—that is, examining response against working memory, suppressing suggested response, activating a unique schema, and substituting and producing novel responses. Fisk and Sharp (2004) and Van der Linden et al. (1998) found effects of age whereby older adults were relatively impaired in their ability to generate random numbers and suggested that RIG is related to both inhibition and the updating component of executive function. A similar pattern was observed when healthy older adults generated random nouns. Categorical recurrence and inhibition have also been shown to respectively increase and decrease with age (Heuer, Janczyk, & Kunde, 2010), again suggesting limited inhibitory processing capacity in older adults. This was confirmed in our study.

Our findings also show a clear positive relationship between the number of symbols uttered and semantic fluency—individuals who scored higher on the semantic fluency task also produced longer sequences in the time frame. The relation of RIG to semantic fluency may be related to roles of selective attention and working memory, as was previously demonstrated in individuals with Alzheimer’s disease (Brugger et al., 1996). In that study, the authors showed that individuals with Alzheimer’s disease who scored lower on semantic fluency tests also generated less evenly distributed random number sequences. Indeed, semantic fluency can be affected by executive function, which may be mediated by skills such as working memory and inhibition (Kiefer, Ahlegian, & Spitzer, 2005).

As a final point, List Recognition scores, which correspond to recall of a previously learned list, correlate positively with the entropy, and negatively with the skewness, of bigraphs in RIG. In other words, a person’s ability to recall lists is associated with increasing uniformity over pairs of symbols in RIG and with decreased clustering over those pairs. It seems likely that the cognitive mechanisms responsible for retaining lists also support the generation of bigraph sequences, and that speakers evaluate the novelty of these sequences during RIG performance. We intend to explore this notion further in the context of inhibitory deficits, with additional data over a larger sample.

A limitation of the study remains the relatively small amount of data used for this task. We are currently exploring much wider-scale data collection, which includes many of the subtasks evaluated here, in an online tool called Talk2Me (https://www.cs.toronto.edu/talk2me). This will allow for more complete statistical comparisons and will also involve a modest cohort of individuals with Alzheimer’s disease. A second limitation is that without direct neuroimaging data, our inferences regarding executive function are constrained to behavioral measures.

Our version of the RIG task provides a fast and repeatable paradigm to infer features relevant to executive function and selective attention. We are currently considering further modifications to make it more practical for clinical trials. Using a machine-learning paradigm with statistical measures, including RQA, we have demonstrated that RIG is reliably related to specific functions assessed by standard relatively complex clinical tools. For example, our neural-network approach can estimate the Stroop interference effect with 1.0% relative error. These results are preliminary but significant and suggest a promising potential in applying machine learning to inferring multivariate measures of cognitive function.

Acknowledgments

This research was undertaken, in part, thanks to funding from the Canada Research Chairs program, awarded to Pascal van Lieshout. Frank Rudzicz was supported by a startup grant from the Toronto Rehabilitation Institute–University Health Network, Natural Sciences and Engineering Research Council Discovery Grant RGPIN 435874, and a Young Investigator award from the Alzheimer’s Society of Canada. We thank Francesca Granata for her suggested changes to the final proof.

References

