

1

2

3

4 **A genetic algorithm to find optimal reading test word**
5 **subsets for estimating full-scale IQ**

6

7

8 Ian van der Linde^{1,2*}, Peter Bright^{2,3}

9

10

11 ¹Department of Computing & Technology, Anglia Ruskin University, Cambridge CB1

12 1PT, United Kingdom

13 ²Vision & Eye Research Unit (VERU), School of Medicine, Anglia Ruskin University,

14 Cambridge CB1 1PT, United Kingdom

15 ³Department of Psychology, Anglia Ruskin University, Cambridge CB1 1PT, United

16 Kingdom

17

18 * Corresponding author

19 E-mail: ian.vanderlinde@anglia.ac.uk (IVDL)

20 **Abstract**

21 In clinical neuropsychology the cognitive abilities of neurological patients are
22 commonly estimated using well-established paper-based tests. Typically, scores on
23 some tests remain relatively well preserved, whilst others exhibit a significant and
24 disproportionate decline. Scores on those tests that measure preserved cognitive
25 functions (so-called ‘hold’ tests) may be used to estimate premorbid abilities,
26 including scores in non-hold tests that would have been expected prior to the onset
27 of cognitive impairment. Many hold tests entail word reading, with each word being
28 graded as correctly or incorrectly pronounced. Inevitably, such tests are likely to
29 contain words that provide little or no diagnostic power (i.e., can be eliminated
30 without negatively affecting prediction accuracy). In this paper, a genetic algorithm
31 is developed and demonstrated, using $n = 92$ neurologically healthy participants, to
32 identify optimal word subsets from the National Adult Reading Test that minimize
33 the mean error in predicting the most widely used clinical measure of IQ and
34 cognitive ability, the Wechsler Adult Intelligence Scale Fourth Edition IQ. In addition
35 to requiring only 17 – 20 of the original 50 words (suggesting that this test could be
36 revised to be up to 66% shorter) and minimizing mean prediction error, the
37 algorithm increases the proportion of the variance in the predicted variable
38 explained in comparison to using all words (from $r^2 = 0.46$ to $r^2 = 0.61$). In a
39 clinical setting this would improve estimates of premorbid cognitive function and, if
40 an abbreviated revision to this test were to be adopted, reduce the arduousness of
41 the test for patients. The proposed method is evaluated with jackknifing and leave
42 one out cross validation. The general approach may be used to optimize the

relationship between any two psychological tests by finding the question subset in one test that minimizes the prediction error in a second test by training the genetic algorithm using data collected from participants upon whom both tests have been administered. This approach may also be used to develop new predictive tests, since it provides a method to identify an optimal subset of a set of candidate questions (for which empirical data have been collected) that maximizes prediction accuracy and the proportion of variance in the predicted variable that can be explained.

50

51 **Introduction**

52 A 'hold test' is a neuropsychological test that measures cognitive functions
53 that remain relatively well preserved following neurological damage caused by
54 traumatic brain injury, stroke, dementia or other condition. In longitudinal studies of
55 preclinical to clinical populations, the relative preservation of hold test performance
56 has been convincingly demonstrated [1]. Since, in neurologically healthy populations,
57 performance in hold tests is highly correlated with that in non-hold tests [2], hold
58 tests can be used with clinical populations to infer premorbid cognitive ability, such
59 as full-scale IQ on the Wechsler Adult Intelligence Scale (WAIS-IV; [3]; for discussion
60 see [4]). Knowledge of premorbid cognitive ability is essential both in evaluating the
61 severity of impairment and in treatment planning.

62 Examples of hold tests that involve reading include the National Adult
63 Reading Test (NART; [5-6]) and its international derivatives (which include NAART
64 and AMNART [USA], [7-9]; NART-SWE [Sweden], [10]; NZART [New Zealand], [11,12];
65 fNART [France], [13]; DART [Netherlands], [14]; and AUSNART [Australia], [15], the

66 Wechsler Test of Adult Reading (WTAR; [16]), the Test of Premorbid Functioning
67 (TOPF; [17]), and a component of the Wide Range Achievement Test (WRAT4; [18]).
68 Although the TOPF is intended to supersede the WTAR, the WTAR is still widely used
69 and the NART also remains popular [19-21], particularly for research purposes.

70 To develop new neuropsychological tests, and to explore the relationships
71 between those already in use, data from multiple tests are collected from healthy
72 participants. In this way, the ability of hold-tests to predict the most likely results in
73 other tests (such as full-scale IQ) can be evaluated (although subsequent longitudinal
74 validation with preclinical to clinical populations is also desirable). In existing studies,
75 a linear regression equation relating reading test performance to full-scale IQ is
76 typically calculated (e.g., [12,20,21]. Ideally, a hold test would yield a perfectly linear
77 correlation with a non-hold test of interest ($r = \pm 1$) and produce perfectly accurate
78 predictions; however, in practice, this goal is unrealistic due to inherent limitations
79 in test reliability and the imperfectly linear relationship expected between any two
80 empirical datasets, especially when they measure different (albeit highly correlated)
81 cognitive functions. The wealth of expertise and normative data relating to existing
82 reading tests means that modifications either to the test or its corresponding
83 instructions are undesirable without compelling justification. However, it is possible
84 to use optimization and artificial intelligence (AI) techniques to develop new tests or
85 revisions to existing tests that are demonstrably superior, or to identify more
86 effective scoring procedures that may be applied to standard tests, e.g., by using
87 question weighting schemes or question subsets that minimize the error between
88 prediction and measurement with experimental data collected from participants
89 upon whom both tests have been administered. In one recent study [22], a genetic

90 algorithm (GA; [23]) was used to produce an abbreviated form of the Psychopathic
91 Personality Inventory – Revised (PPI-R). In another, a GA was used to abbreviate the
92 Multidimensional Experiential Avoidance Questionnaire [24]. Similarly, a GA with
93 logistic regression to select the optimum combination of neuropsychological test
94 results to predict progression to Alzheimer’s disease [25]. In the present study,
95 rather than using genetic algorithms to abbreviate a test for comparison against
96 results obtained using the full test, we use a GA to identify the optimum question
97 subset from one test to most accurately estimate the result of a second (predicted)
98 test.

99 A related area of research abbreviates tests on a per-participant basis. In
100 Computerized-adaptive Testing (CAT) questions are selected based upon an estimate
101 of current performance, and can yield accuracy comparable to an equivalent full-
102 length test in which all questions are used [26]. In Multi-stage Testing (MST; [27]) a
103 broadly similar approach is taken, except that banks of questions (so-called *testlets*)
104 are selected at each decision stage. Using these approaches, sequences of decisions
105 are made on-the-fly concerning which questions to present. However, such
106 approaches are not appropriate in this case, where a core subset of questions is to
107 be developed from which a single linear regression equation is desired, for which
108 tests are administered by the clinician on paper (rather than using a computer).
109 Additionally, the standardized instructions for the NART, used in this article to
110 illustrate the general approach, require all items to be attempted for scoring to be
111 valid. Furthermore, the approach presented is well suited to test design, enabling
112 the researcher to develop new tests by establishing optimum combinations of

113 questions that maximize predictive accuracy, potentially based upon a parent test
114 (such as the NART).

115 To illustrate the general approach, data from the British NART [5,6] is used, in
116 part because a recent survey indicates that is the most widely cited [21], but also
117 because has been made freely available for use without restriction. It comprises 50
118 visually presented words that have irregular non-phonetic spellings and for which
119 verbal responses elicited from participants are subject (by the experimenter,
120 following standardized instructions) to binary classification as either having been
121 correctly or incorrectly pronounced. The NART is scored by counting the number of
122 incorrectly pronounced words (hereafter referred to as NART errors), and the
123 instructions require that participants attempt all words for the scoring to be valid.
124 The irregular nature of the words (i.e., their violation of typical phoneme-grapheme
125 correspondence rules) is such that participants should be unable to spontaneously
126 deduce correct pronunciations, and as such the test measures prior knowledge [28].
127 The set of 50 words that feature in the NART generally increase in difficulty through
128 the test (thus the order that the words are presented is fixed, with words presented
129 towards the end of the test intended to be less familiar to the target population). A
130 patient who has suffered neurological impairment may therefore find the test rather
131 onerous, particularly towards the end when presented with a sequence of
132 increasingly difficult words. Furthermore, the intentionally ramped difficulty may
133 disproportionately affect particular patient types for whom increased fatigue and
134 impairments in concentration are apparent, making the use of an abbreviated test
135 both faster to administer and less susceptible to confounds arising from patient
136 fatigue.

137 At present, to predict premorbid intelligence using the NART, a linear
138 regression equation is calculated in which the explanatory variable is NART errors
139 and the predicted variable is, in the most recent standardization [20], WAIS-IV Full-
140 scale IQ (FSIQ). A negative correlation ($r < 0$) is expected, such that an increase in
141 the number of NART errors should yield commensurate reduction in predicted WAIS-
142 IV FSIQ. In this paper, a GA is presented that increases the association between the
143 NART and WAIS-IV FSIQ, reduces mean absolute prediction error, and reduces the
144 number of words that participants are asked to pronounce. This approach is
145 assessed for stability and overfitting via jackknifing [29,30] and exhaustive leave-one
146 out cross-validation [31].

147 In recognition of the possibility that some NART words may provide little or
148 no diagnostic power, and acknowledging that reduced test duration is desirable, in a
149 recent study by McGrory and colleagues [32], Mokken scaling [33-34] was used to
150 produce a reduced (and thus faster to administer) 23-word version of the NART.
151 Referred to as the mini-NART, it was found to account for a similar proportion of
152 variance in FSIQ as the full NART (44.8% vs. 46.5%). In this article, a markedly
153 different approach is used that has several empirical advantages over the use of the
154 full NART or the mini-NART: 1. it accounts for a greater proportion of the variance in
155 measured WAIS-IV FSIQ; 2. residuals between predicted and measured WAIS-IV FSIQ
156 using the identified NART subset are verified to be less than or equal to those
157 observed using the full NART; 3. The number of words that participants are asked to
158 pronounce is reduced from 50 (or 23 for the mini-NART) to 17-20 (around two thirds
159 of the full test), suggesting that the test could be shortened, thereby reducing the
160 likelihood of unnecessary fatigue. Furthermore, the method proposed simply

161 requires the exclusion of individual NART words and the application of a revised
162 regression equation, and can therefore either be administered as an abbreviated
163 test or be applied retrospectively to existing data by rescoreing the identified subset
164 of words. The technique can, more generally, be used in the design of new predictive
165 tests to identify an optimal subset of a set of candidate questions that yields the
166 greatest coefficient of determination and smallest mean residual in relation to the
167 measure that the test is intended to predict.

168

169 **Initial model**

170 **Participants**

171 An opportunity sample of 100 neurologically healthy adults were recruited
172 primarily from University campuses in Cambridge and London, local retail outlets,
173 and via social media, of which eight participants failed to complete one or more tests
174 and were excluded from all analyses. There were no missing data across the sample
175 of 92 participants (mean age 40 years; range 18 – 70; s_{age} 16.78), of which 30 were
176 male, on any of the tests reported here. All were British nationals, with English as
177 the first language, and with normal/corrected-to-normal vision and hearing.
178 Participants self-declared that they had no history of neurological or psychiatric
179 disorder. Extensive training in the administration and scoring of all tests was
180 provided to three research assistants over several days by PB (an experienced
181 neuropsychologist), and the testing sessions were closely monitored and supervised
182 to ensure full compliance with the standardized administration and scoring
183 procedures. All participants were recruited and tested between 2013 and 2016, in a

184 UK University setting. The procedure was approved by the University Ethics Panel,
185 and was conducted in accordance with the tenets of the Declaration of Helsinki. All
186 participants had normal/corrected-to-normal vision and hearing (self reported), and
187 spoke English as their first language.

188

189 **Data collection**

190 All participants completed the NART first and then all 10 core subtests from
191 the WAIS-IV battery. All tests were administered following standard published
192 instructions. Participants attended a single session of approx. 90 minutes, with
193 breaks provided upon request.

194

195 **Analysis procedure and results**

196 The NART responses for each participant were placed in a 2-D bit matrix, to
197 be denoted Q , in which each row ($1..m$) corresponded to a NART word index, and
198 each column ($1..n$) to a participant number (Fig 1). Here, rows $m = 50$ and
199 columns $n = 92$.

200

201 << Figure 1 About Here >>

202

203 **Fig 1. 2-D bit matrix for all participants and NART word responses in which a black**
204 **dot denotes a pronunciation error.**

205

206 The presence of a 1 in Q (a black dot in Fig 1) denotes an incorrect
 207 pronunciation (error), so the total number of NART errors, x_j , for each participant j
 208 from 1.. n over the sequence of NART words i from 1.. m , is given by Eq. 1, such that
 209 $x_j \in [0..50]$. The number of NART errors per participant in our data ranged from 2
 210 to 46 ($\bar{x} = 18.25, s_x = 8.91$). Corresponding WAIS-IV FSIQ results, to be denoted y ,
 211 ranged from 80 to 150 ($\bar{y} = 108.52, s_y = 12.71$). A Kolmogorov-Smirnov test
 212 indicates that neither empirical dataset deviates significantly from a normal
 213 distribution ($k = 0.98, k = 1.00$, both $p < .0001$).

214

$$x_j = \sum_{i=1}^m Q_{i,j}, \quad j = 1, \dots, n \quad (1)$$

215

216 The linear Pearson product-moment correlation coefficient (PPMCC)
 217 between NART errors (x) and measured WAIS-IV FSIQ (y) is given by Eq. 2. In
 218 addition to enabling a linear regression equation to be calculated (see below), the
 219 PPMCC, r , and coefficient of determination, r^2 , are commonly used in
 220 neuropsychological literature to assess the degree of association between different
 221 test scores (e.g., see [28]), and provide one metric against which the GA-derived
 222 model described later in this article is to be evaluated. The value given by Eq. 2 for
 223 our data, consistent with that reported in [20], was $r_{(90)} = -0.68, p < 0.000001$,
 224 which is typically classified as *large* [35]. The coefficient of determination was $r^2 =$
 225 0.47, a comparable number to that reported in [5,6] and many subsequent studies
 226 that correlate NART error scores against earlier iterations of WAIS IQ. It suggests that

the explanatory variable (NART errors) accounts for 47% of the variance in the predicted variable (WAIS-IV FSIQ).

$$r_{xy} = \frac{1}{ns_x s_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (2)$$

A linear regression equation (of the form $\hat{y} = ax + b$, where \hat{y} denotes a predicted value of y) was produced, again in keeping with earlier approaches, with multiplicative constant a (slope) and additive constant b (y -intercept), which can then be used to predict WAIS-IV FSIQ (\hat{y}) for any number of NART errors (x). The PPMCC, r (Eq. 2), is used to calculate the line equation constants (Eq. 3 for slope, a , and then Eq. 4 for y -intercept, b).

$$a = r \frac{s_y}{s_x} \quad (3)$$

$$b = \bar{y} - a\bar{x} \quad (4)$$

Using the full set of NART words, the line equation for our data was $\hat{y} = -0.9750x + 126.3163$, shown on a scatterplot of raw NART errors vs. WAIS-IV FSIQ in Fig 2 as a dotted black line with circles denoting measured values (i.e., our 92 participant test scores). The proximity of the sample points to the initial line equation is highlighted as a shaded zone (convex hull [36]).

<< Figure 2 About Here >>

247

248 **Fig 2. Scatterplot of raw NART errors vs. measured WAIS-IV FSIQ (hollow circles).**

249 **Dotted black line is initial line equation; shaded zone is the convex hull.**

250

251 A correlation coefficient (or coefficient of determination) should not, on its
252 own, be used to assess the accuracy of a linear regression model such as this, since
253 in a comparison between two hypothetical models, greater absolute r (or greater
254 r^2) for the first model may coincide with greater predictive accuracy for the second
255 model, since the slope of a regression line is not necessarily coupled with lower
256 average residuals (i.e., shorter average distance of measured sample points to their
257 corresponding predictions). An additional metric should be used that specifically
258 assesses the accuracy with which a model predicts known values; one simple metric
259 that can accomplish this is mean absolute error (MAE, Eq. 5), which has the
260 advantage of being in the same units as the predicted variable (here, IQ points).
261 Using raw NART errors, $MAE = 7.33$ ($s = 5.64$), showing that, on average, the
262 error between predicted and observed WAIS-IV FSIQ using raw NART errors for our
263 data was 7.33 IQ points.

264

$$MAE = \frac{1}{n} \sum_{j=1}^n |\hat{y}_j - y_j| \quad (5)$$

265

266 In addition, regression models should be validated to examine their stability
267 to the removal of data points (i.e., the degree to which they may be affected by
268 outliers), and their ability to make accurate predictions for samples not used in their

269 production (i.e., the degree to which overfitting may have occurred). Alternative
 270 approaches to accomplish this include dividing data into training and testing sets, k -
 271 folds validation [37], and exhaustive leave-one-out cross-validation (LOOCV),
 272 described in [31]. The latter approach is used here, in part because it is fully
 273 reproducible (i.e., does not depend upon the randomized division of data into
 274 training and testing subsamples). In this form of validation, the predicted variable
 275 and other metrics of interest are calculated using models produced using
 276 subsamples of the original data in which one participant at a time has been left out
 277 (i.e., n subsamples of $n - 1$ participants, with participant k left out, such that k is
 278 iterated from $1..n$). These are sometimes referred to as jackknife samples.
 279 Thereafter, the accuracy with which each of the n models predict metrics of interest
 280 for the one left out participant not used their production is assessed. As before,
 281 MAE may be used to evaluate prediction accuracy both for the n jackknife models
 282 (which comprised $n(n - 1) = 8372$ individual predictions) and the n single left out
 283 sample predictions (here 92). A correlation coefficient (or coefficient of
 284 determination) can only be produced for the jackknife models since the left out
 285 samples are not associated with a single line equation.

286 For our data, averaging over the n jackknife models, with standard deviation
 287 shown in parenthesis, yields $r = -0.68$ (0.01), $a = -0.9750$ (0.01), $b =$
 288 126.3157 (0.28), and $MAE = 7.33$ (5.61). These values are remarkably close to
 289 where all participant data were used (reported above), indicating that outliers did
 290 not significantly affect these metrics. Next, the accuracy of the predicted variable
 291 elicited by each model using each respective single left-out participant (i.e., the
 292 participant not used in the production of that model) was assessed. This yielded

293 $MAE = 7.49$ ($SD = 5.78$), which is fractionally greater than the MAE calculated
294 using all data and the average MAE across the n jackknife models; however, this is
295 to be expected given that the models are now being requested to make predictions
296 for participants that were not used in their production. Furthermore, the differences
297 in MAE values (between all data, 8372 jackknife subsamples, n leave-one-out
298 samples) were not statistically significant ($p > .05$).

299

300 **Genetic algorithm model**

301 **Apparatus**

302 Statistical analyses and optimization algorithms were implemented in
303 MATLAB (The Mathworks Inc., Natick MA). The standard regression mode, GA,
304 validation routines and experimental data used for testing and validation are freely
305 provided for download from the Open Science Framework
306 (<http://dx.doi.org/10.17605/OSF.IO/34BKU>).

307

308 **Analysis procedure**

309 The GA described below is charged with finding the optimum subset of NART
310 words that yields the smallest average prediction residual (MAE), working from the
311 initial starting point of using all 50 words.

312 GAs search solution spaces so large that they cannot feasibly be traversed
313 using exhaustive/analytical approaches, enabling them to address computational
314 problems, like the present one, that have no polynomial-time exhaustive solution.
315 The final solution returned by a GA is not necessarily the best possible answer, since

316 they rely upon an adaptive heuristic approach that iteratively improves upon each
317 currently held solution until a solution that is deemed acceptably good is obtained.
318 However, if appropriately configured, GAs can produce solutions that dramatically
319 improve upon the initial starting point. GAs, being inspired by the biological principle
320 of natural selection by survival of the fittest, entail the representation of candidate
321 solutions as *chromosomes*, the evaluation of chromosome efficacy through a *fitness*
322 *function*, the creation of new chromosomes via *mutation* and/or *crossover*
323 (principally from the chromosomes identified as the most fit), and a *selection*
324 method by which individual chromosomes are chosen to sire subsequent
325 generations. A *termination* criterion must also be decided upon to determine how
326 long the GA will run. Alternatives include letting the GA run for a fix period of time,
327 for a fixed number of generations, until the solution is valid (e.g., in some NP-class
328 problems in which finding a solution that merely works is a laudable goal), or until
329 the fitness of the solutions produced over a pre-determined period of time or
330 number of generations ceases to improve (i.e., evolutionary stagnation).

331

332 **Chromosome structure**

333 Each chromosome, c , was a 1-D bit string (specifically, a sequence of 50
334 binary digits, each referred to as a gene) wherein each bit controls whether the
335 NART word at index i should be used ($c_i = 1$) or not used ($c_i = 0$) in the calculation
336 of each participant's revised NART score. The number of alleles (alternatives) for
337 each gene was therefore 2: 0 and 1. All possible solutions to the problem of finding
338 the optimum NART subset can be represented on such a chromosome, of which

339 there are 2^{50} (one quadrillion, one hundred twenty five trillion, eight hundred ninety
340 nine billion, nine hundred six million, eight hundred forty two thousand, six hundred
341 and twenty four), which is the cardinality of the powerset of the set of words (w) in
342 the original NART, $|\wp(w)| = 2^{|w|}$ (i.e., the size of the set of all possible subsets of w).
343 If one were to iterate through these subsets one at a time, a tight bound algorithm
344 of exponential time complexity $\Theta(2^{|w|})$ would be required, which is computationally
345 impractical. To put this into perspective, if each subset took 1 sec to evaluate, it
346 would take 36 million years to sequentially test all subsets to identify the true
347 (guaranteed) optimum.

348

349 **Settings**

350 The GA was run for a fixed number of generations (128), which was found to
351 be more than sufficient for fitness to reach a stable asymptote (i.e., after which no
352 further improvement in fitness was observed), and also provided an acceptable run-
353 time of approx. 1-2 minutes on a standard computer. The number of children
354 produced in each generation was also set to 128. With these settings, one run
355 produces a total of $128 \times 128 = 16384$ candidate solutions, with a tendency to
356 improvement from generation-to-generation that inevitably slows as the algorithm
357 progresses and fitter solutions become more difficult to find. A mutation rate [38] of
358 $\frac{5}{50}$, i.e. 10%, was used, determined experimentally to, when coupled with the use of
359 128 children per generation, yield fast and stable evolutionary descent.

360 The mutation routine entailed the negation of randomly selected genes
361 (sometimes called *bit mutation* or *bit flipping*). Crossover (recombination) was not

362 used, since this is not thought to be an effective approach for problems in which
 363 large changes in chromosome composition are likely to dramatically affect
 364 performance and thereby thwart evolutionary progress. A simple maximally elitist
 365 GA was used, such that only the fittest child in each generation was retained
 366 (determined as described below), which was set to be the parent of the subsequent
 367 generation. Other (less fit) chromosomes were destroyed. However, the fittest child
 368 in each generation always replaced the parent chromosome (whether fitter or not),
 369 enabling the fitness profile over time to decrease as well as increase, which is
 370 thought to prevent premature convergence (although it is acknowledged that the
 371 single-parent approach could lead to convergence to local optima, to demonstrate
 372 the general procedure, this simple approach was taken, and suitably fit solutions
 373 were indeed produced).

374

375 **Fitness function**

376 To calculate chromosome fitness (a so-called *figure of merit*), first a revised
 377 NART response matrix, Q' , is calculated from Q , the original response matrix, by
 378 multiplying each participant's NART word responses with the chromosome to be
 379 evaluated (Eq. 6). This had the effect of masking the responses for specific words so
 380 that they no longer contributed to the final score for any participant.

381

$$Q'_{i,j} = \sum_{i=1}^m Q_{i,j} \cdot c_i, \quad j = 1..n \quad (6)$$

382

Next, the number of NART errors for the surviving NART words only, to be called the revised NART score, x' , is calculated using Eq. 1 with Q' substituting for Q . Next, a correlation coefficient is calculated using Eq. 2 with x' substituting for x , and then revised line equation constants, a' and b' , are calculated using Eqs. 3 and 4 with \bar{x}' and $s_{x'}$ substituting for \bar{x} and s_x , respectively. A revised prediction, \hat{y}' can then be calculated for each participant. Using Eq. 5, with \hat{y}' substituting for \hat{y} , the *MAE* using the adjusted NART scores can then be calculated, evaluating how accurately the current word subset approximates measured WAIS-IV FSIQ. The *MAE* value is returned as the fitness of the chromosome.

Over successive generations, the GA identifies the optimum subset: i.e., the optimum values in c , that when entrywise multiplied with the raw NART responses from all participants identically, minimizes *MAE*. As a consequence of falling *MAE*, the absolute correlation and coefficient of determination will also typically increase from generation-to-generation. It is worth noting that there may be multiple equally fit chromosomes, and that running the GA on multiple occasions could produce subtly different word subsets each time because several words or word combinations may each be equally suitable alternatives in c for *MAE* minimization (reflecting the randomness inherent in evolutionary descent, akin to nature). Selecting for the additional criterion of minimal subset cardinality is one approach that could be used to select from among equally fit alternatives (i.e., when two subsets yielding equal *MAE* are evaluated, particularly if the objective was to devise an abbreviated test or a new test using the fewest number of candidate questions).

405

406 **Model results**

7.32 (6.11), with a mean prediction of 108.69 (9.79). The critical performance metrics (subset cardinality, mean prediction, r , and r^2 , and MAE) are shown in relation to the original model and its validation in Table 1. It is apparent that the cardinality of the subset (number of retained NART words) is fractionally higher than the ‘best’ runs using all data described above (in which 17 words are retained). Indeed, due to the heuristic nature of the approach, running the all-data single-run GA multiple times also produces some results wherein 19-20 words are retained, since these alternative solutions also yield an $MAE = 5.75$. It likely that, with if the number of participants were increased, MAE for single left-out participants would fall.

<< Figure 5 About Here >>

Fig 5. Chromosomes for n jackknifed GA models in which participant k was left out.

In Table 1, it is apparent that the MAE for the all-data GA model and its jackknife subsamples are lower than the initial model all-data MAE and its jackknife subsamples. Comparing jackknife distributions using a t -test, this difference, although relatively small, is statistically significant [$t_{(8462)} = 2.75, p < 0.01$]. The MAE differences between the one-left-out (validation) sets in the initial and GA-derived models is not statistically significant ($p = 0.84$), despite that the GA-derived model uses, on average, only 19-20 words of the original 50, demonstrating that the additional words in the NART, as originally formulated, did not improve predictive accuracy for our data.

478

479 **Table 1. Critical performance metrics for each models and cross validation set**
 480 **(standard deviation shown in parenthesis, were available).**

	Initial Model	Cross Validation		GA Model (Best of 100)	GA Cross Validation	
		Jackknife	One Left Out		Jackknife	One Left Out
Cardinality (Words)	50	50	50	17	19.65 (0.80)	19.65 (0.80)
Mean Prediction	108.52 (12.71)	108.52 (8.64)	108.53 (8.66)	108.52 (9.94)	108.52 (9.85)	108.69 (9.79)
PPMCC (r)	-0.68	0.68 (0.01)	-	-0.78	-0.78 (0.01)	-
COD (r²)	0.46	0.46	-	0.61	0.61	-
MAE	7.33 (5.64)	7.33 (5.61)	7.49 (5.78)	5.75	5.76 (5.45)	7.32 (6.11)

481
482

483

484 **Conclusions**

485 A GA for optimizing the relationship between neuropsychological test data is
 486 presented and demonstrated using the NART and WAIS-IV FSIQ leading to increased
 487 absolute correlation/coefficient of determination, potentially reduced mean
 488 absolute error (i.e., smaller prediction residuals). The GA suggests that the number
 489 of words in the NART may be reduced by up to 66%; however, to evaluate the
 490 potential effects of reduced fatigue and alternate word order, the use of an optimal
 491 word subset should ultimately be evaluated by collecting data using it directly,
 492 rather than having data corresponding to the words identified as having predictive
 493 value being extracted from an administration of the full test. Due to the way that the
 494 GA was implemented, it could be that evolution to a local optimum occurred, and
 495 that a global optimum with higher performance is possible. It is also possible that
 496 incorporating other information, such as participant demographics and results on
 497 other hold tests (such as that WTAR and TOPF), may further elevate the correlation

498 and reduce mean prediction error, rather like Johnson et al. [25] who used a GA to
499 find the best combination of different neuropsychological tests to predict
500 progression to Alzheimer's disease (except that, here, questions from within tests
501 would be selected, rather than tests themselves). Furthermore, greater performance
502 might be achieved using artificial neural networks or other alternative AI approaches,
503 or by weighting individual words rather than by the creation of optimal subsets.
504 These possibilities are under investigation; however, the current article serves to
505 illustrate a general principle that the strength of association between
506 neuropsychological tests may be increased using GAs by, in this article, selecting
507 optimum question subsets. A further caveat is that the cross-validation routines
508 used here, although including LOOCV in which models are tested upon individual
509 samples upon which they were not trained, may still be susceptible to a degree of
510 overfitting; this possibility can be investigated in a follow-up study with a larger
511 cohort which would support the division of participants into adequately sized
512 training and validation sets.

513 It may also be the case that different locations (e.g., clinical centers, or
514 geographic regions), participant demographics, and other clinical indicators may
515 influence the optimal subset, so it may be more effective to select data to determine
516 the optimal subset using one or more of these parameters, all without needing to
517 adjust the basic NART test procedure, retaining the simplicity of administering this
518 test for clinicians and enabling it to be used retrospectively on data already collected.
519

520 **References**

- 521 1. McGurn B, Starr JM, Topfer JA, Pattie A, Whiteman MC, Lemmon HA, Whalley LJ,
522 Deary IJ. Pronunciation of irregular words is preserved in dementia, validating
523 premorbid IQ estimation. *Neurology*. 2004; 62:1184–1186.
- 524 2. Lezak MD. *Neuropsychological assessment*. 5th ed. Oxford University Press:
525 Oxford UK; 2012.
- 526 3. Wechsler D. *Wechsler Adult Intelligence Scale*. 4th ed. San Antonio, TX: Pearson
527 Assessment; 2008.
- 528 4. Lichtenberger EO, Kaufman AS. *Essentials of WAIS-IV assessment*. John Wiley &
529 Sons; 2009.
- 530 5. Nelson HE. *The National Adult Reading Test (NART)*. Windsor: NFER-Nelson; 1982.
- 531 6. Nelson HE, Willison J. *The National Adult Reading Test (NART)*. Windsor: NFER-
532 Nelson; 1991.
- 533 7. Blair JR, Spreen O. Predicting premorbid IQ: a revision of the National Adult
534 Reading Test. *Clin Neuropsychol*. 1989; 3:129-136.
- 535 8. Grober E & Sliwinski M. Development and validation of a model for estimating
536 premorbid verbal intelligence in the elderly. *J Clin Exp Neuropsychol*. 1991;
537 13:933-949.
- 538 9. Gladsjo, JA, Heaton RK, Palmer BW, Taylor MJ, & Jeset, DV. Use of oral reading to
539 estimate premorbid intellectual and neuropsychological functioning. *J Int*
540 *Neuropsychol Soc*. 1999; 5:247-254.
- 541 10. Rolstad, S. The Swedish National Adult Reading Test (NART-SWE): A test of
542 premorbid IQ. *Scand J Psychol*. 2008; 49: 577–582.
- 543 11. Starkey NJ, Halliday T. Development of the New Zealand Adult Reading Test
544 (NZART): Preliminary findings. *NZ J Psychol*. 2011; 40:129- 141.

- 545 12. Lichtwark IT, Starkey NJ, & Barker-Collo S. Further validation of the New Zealand
546 Test of Adult Reading (NZART) as a measure of premorbid IQ in a New Zealand
547 sample. *NZ J Psychol.* 2013; 42:60-68.
- 548 13. Mackinnon A, Ritchie K, Mulligan R. The measurement properties of a French
549 language adaptation of the National Adult Reading Test. *Int J Methods Psychiatr*
550 *Res.* 1999; 8:27-38.
- 551 14. Schmand B, Bakker D, Saan R, Louman J. The Dutch Reading Test for Adults: a
552 measure of premorbid intelligence level. *Tijdschr Gerontol Geriatri.* 1991; 22:15-
553 19.
- 554 15. Hennessy M, Mackenzie B. AUSNART: The development of an Australian version
555 of the NART. In J. Fourez & N. Page (Eds.), *Treatment issues and long-term*
556 *outcomes: Proceedings of the 18th Annual Brain Impairment Conference* (pp. pp.
557 183-188). Bowen Hills: Australian Academic Press; 1995.
- 558 16. Wechsler D. Wechsler Test of Adult Reading: WTAR. Psychological Corporation;
559 2001.
- 560 17. Wechsler D. Test of Premorbid Functioning. UK version (TOPF UK). UK: Pearson
561 Corporation; 2011.
- 562 18. Wilkinson GS, Robertson GJ. Psychological Assessment Resources. In *Wide Range*
563 *Achievement Test 2006. WRAT4 Introductory Kit.*
- 564 19. Bright P, Jaldow EL, Kopelman MD. The National Adult Reading Test as a measure
565 of premorbid intelligence: a comparison with estimates derived from
566 demographic variables. *J Int Neuropsychol Soc.* 2002; 8:847-854.

- 567 20. Bright P, Hale E, Gooch VJ, Myhill T, van der Linde I. The National Adult Reading
568 Test: restandardisation against the Wechsler Adult Intelligence Scale—Fourth
569 edition. *Neuropsychol Rehabil*. 2016. doi: 10.1080/09602011.2016.1231121
- 570 21. Bright P, van der Linde I. Comparison of methods for estimating premorbid
571 intelligence. *Neuropsychol Rehabil*. 2018. doi: 10.1080/09602011.2018.1445650
- 572 22. Eisenbarth H, Lilienfeld SO, & Yarkoni T. Using a genetic algorithm to abbreviate
573 the Psychopathic Personality Inventory—Revised (PPI-R). *Psychol Assess*. 2015;
574 27: 194-202.
- 575 23. Holland JH. *Adaptation in natural and artificial systems*. Ann Arbor, MI: University
576 of Michigan Press; 1975.
- 577 24. Sahdra BK, Ciarrochi J, Parker P, Scrucca L. Using genetic algorithms in a large
578 nationally representative American sample to abbreviate the Multidimensional
579 Experiential Avoidance Questionnaire. *Front Psychol*. 2016; 7: 189.
- 580 25. Johnson P, Vandewater L, Wilson W, Maruff P, Savage G, Graham P, Macaulay LS,
581 Ellis KA, Szoek C, Martins RN and Rowe CC. Genetic algorithm with logistic
582 regression for prediction of progression to Alzheimer's disease. *BMC*
583 *Bioinformatics*. 2004; 15:S11.
- 584 26. Weiss DJ, Kingsbury GG. Application of computerized adaptive testing to
585 educational problems. *J Educ Meas*. 1984; 21:361–375.
- 586 27. Luecht RM, Nungester RJ. Some practical examples of computer-adaptive
587 sequential testing. *J Educ Meas*. 1998; 35:229-249.
- 588 28. Crawford JR, Venneri A, and O'Carroll RE. Neuropsychological assessment of the
589 elderly. In: A. S. Bellack and M. Herson (Eds), *Comprehensive Clinical Psychology*
590 Vol. 7, *Clinical Gerontology* (pp. 133-169). Pergamon, Oxford UK; 1998.

- 591 29. Tukey JW. Bias and confidence in not quite large samples. *Ann Math Stat.* 1958;
592 29: 614–623.
- 593 30. Efron B. The jackknife, the bootstrap and other resampling plans. Society for
594 industrial and applied mathematics. In *CBMS-NSF Regional Conference Series in*
595 *Applied Mathematics*, 1982, Philadelphia, PA: Society for Industrial and Applied
596 *Mathematics*; 1982.
- 597 31. Hastie T, Tibshirani R, and Friedman J. *The elements of statistical learning: Data*
598 *mining, inference and prediction.* 2nd Ed. NY: Springer-Verlag; 2009.
- 599 32. McGrory S, Austin EJ, Shenkin SD, Starr JM, Deary IJ. From “aisle” to “labile”: A
600 hierarchical National Adult Reading Test scale revealed by Mokken scaling.
601 *Psychol Assess.* 2015; 27: 932-943.
- 602 33. Mokken RJ. *A theory and procedure of scale analysis.* de Gruyter, Berlin,
603 Germany; 1971.
- 604 34. Mokken RJ, Lewis C. A nonparametric approach to the analysis of dichotomous
605 responses. *Appl Psychol Meas.* 1982; 6: 417–430.
- 606 35. Cohen J. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale,
607 NJ: Lawrence Erlbaum Associates; 1988.
- 608 36. de Berg M, van Kreveld M, Overmars M, Schwarzkopf O. In *Computational*
609 *Geometry:* (pp. 1-17), Springer Berlin Heidelberg; 2000.
- 610 37. Stone M. Cross-validatory choice and assessment of statistical predictions. *J*
611 *Royal Stat Soc.* 1974; 36:111–147.
- 612 38. Fogel D. *Evolutionary computation: Toward a new philosophy of machine*
613 *intelligence.* 3rd ed. Piscataway, NJ: IEEE Press; 2006.

Figure 1

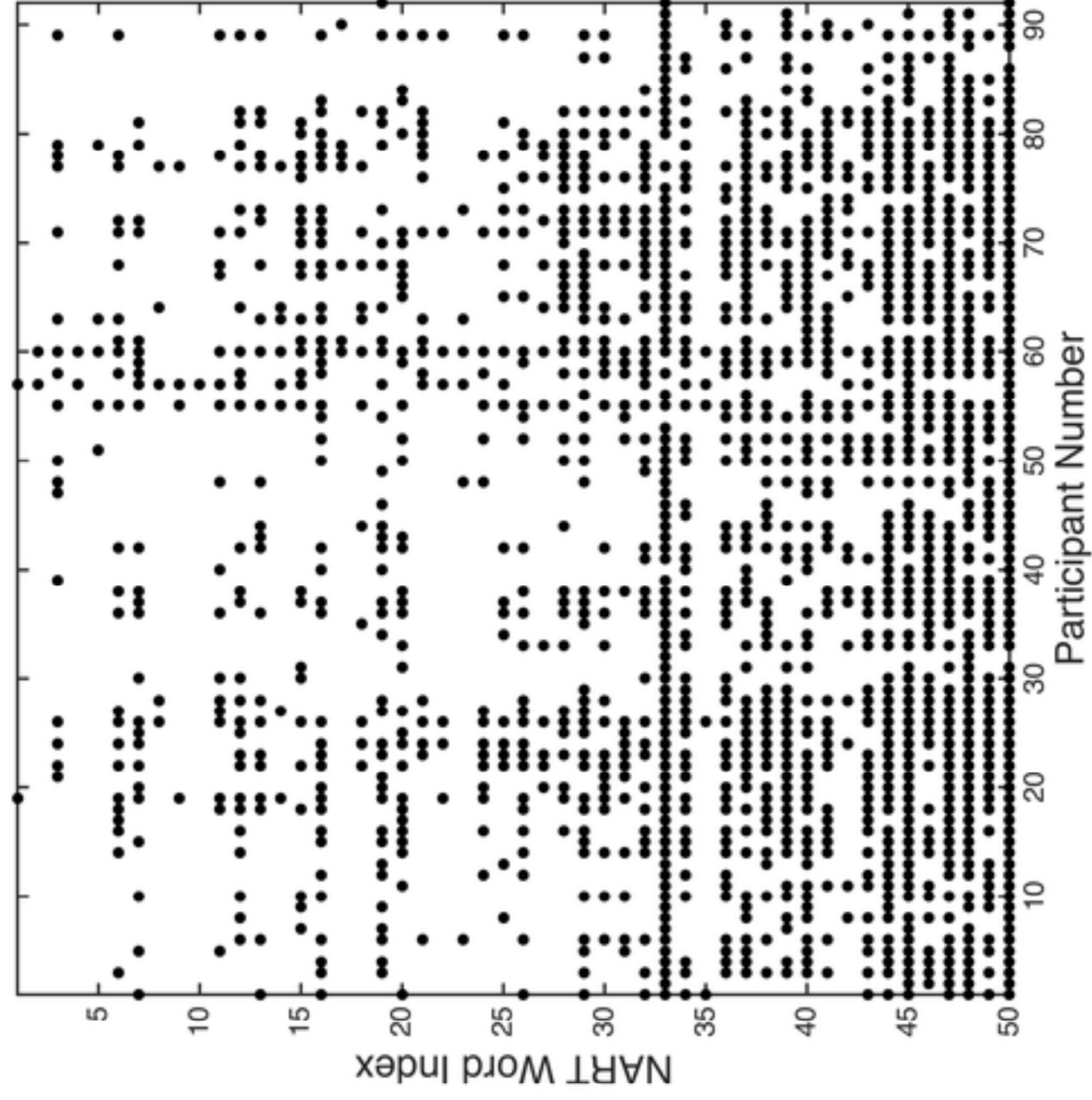


Figure 2

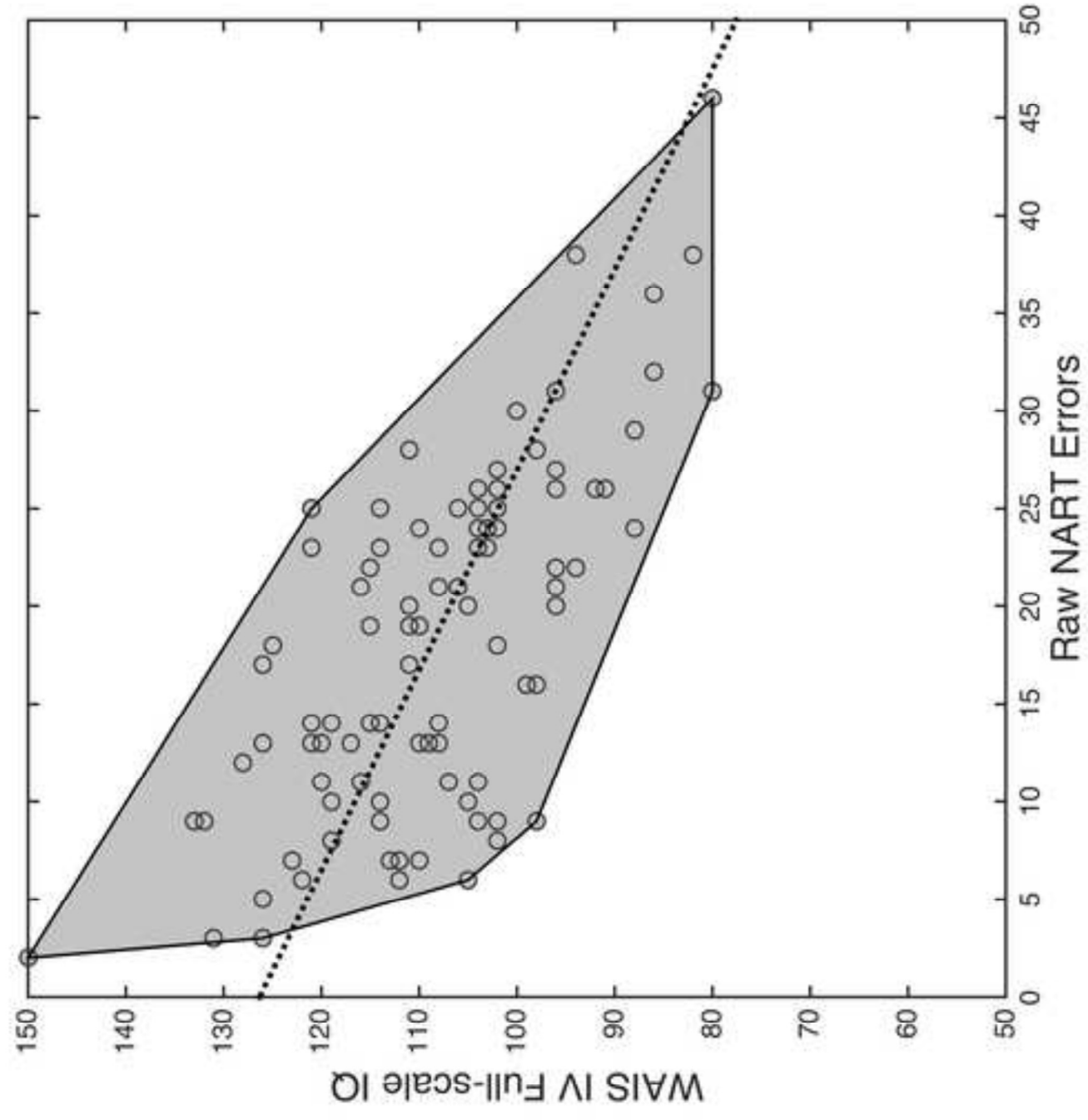


Figure 3

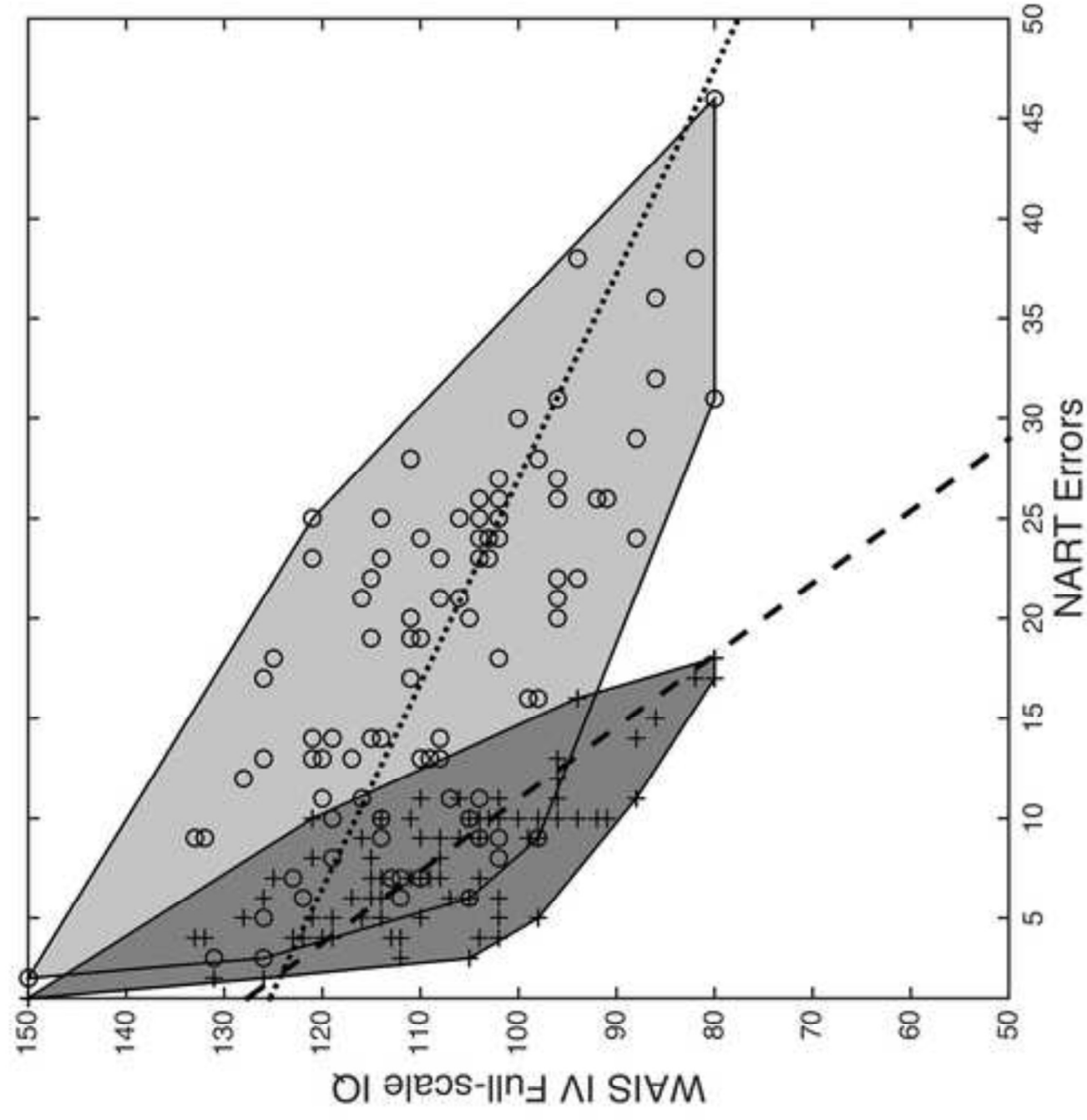


Figure 4

[Click here to access/download;Figure;Figure 4.tiff](#)

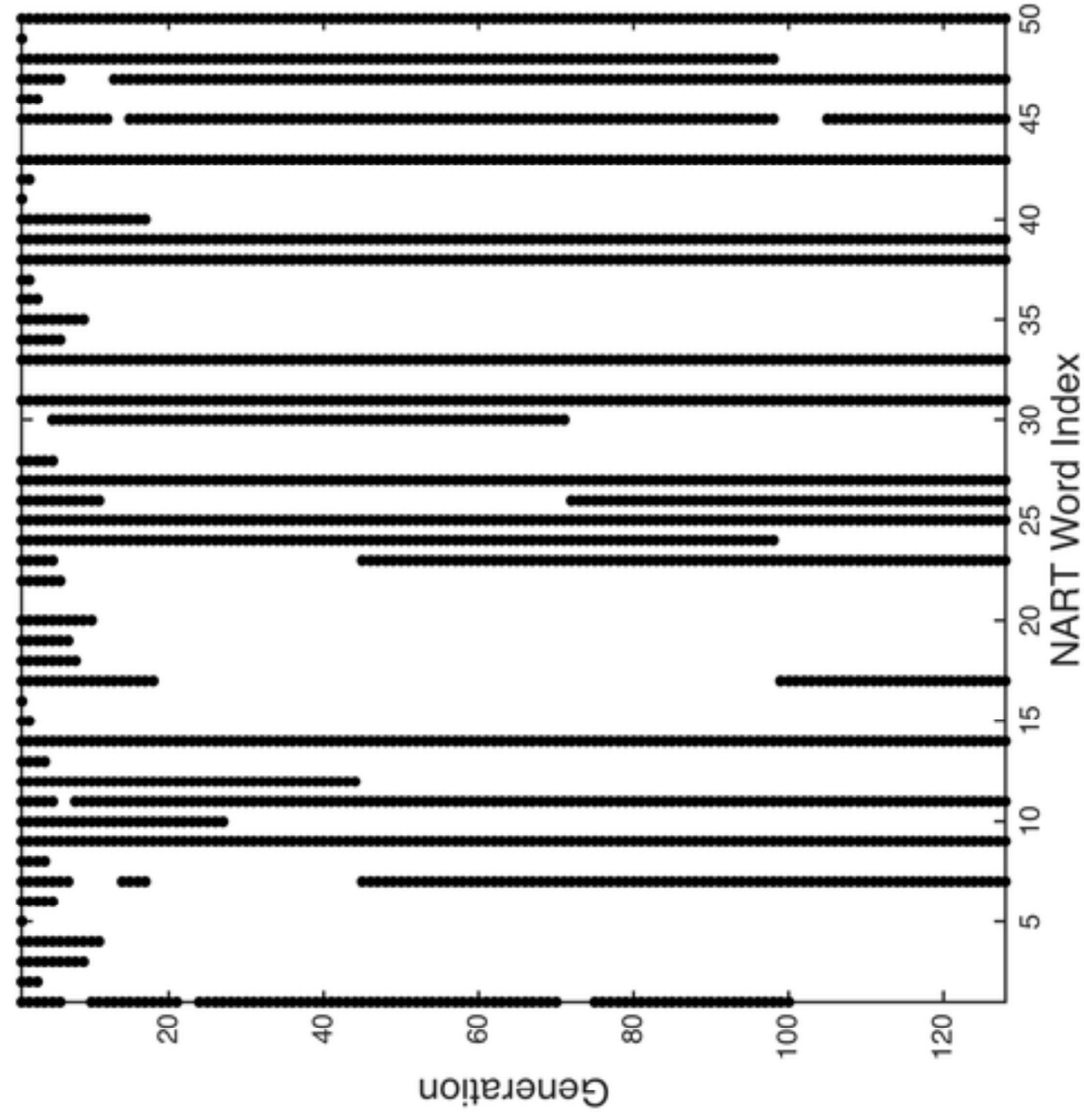


Figure 5

