Published in final edited form as:

Brain Cogn. 2011 June ; 76(1): 43–51. doi:10.1016/j.bandc.2011.03.004.

A Twin Study of Spatial and Non-Spatial Delayed Response Performance in Middle Age

William S. Kremen^{1,2}, Tuan Mai¹, Matthew S. Panizzon¹, Carol E. Franz¹, Howard M. Blankfeld³, Hong Xian⁴, Seth A. Eisen⁵, Ming T. Tsuang^{1,2}, and Michael J. Lyons⁶

¹ Department of Psychiatry, University of California, San Diego, La Jolla, CA

² Center for Behavioral Genomics, University of California, San Diego, La Jolla, CA

³ Private Practice, Lafayette, CA

⁴ St. Louis VA Medical Center and Department of Internal Medicine, Washington University, St. Louis, MO

⁵ Department of Veterans Affairs, Washington, DC and Departments of Medicine and Psychiatry, Washington University, St. Louis, MO

⁶ Department of Psychology, Boston University

Abstract

Delayed alternation and object alternation are classic spatial and non-spatial delayed response tasks. We tested 632 middle-aged male veteran twins on variants of these tasks in order to compare test difficulty, measure their inter-correlation, test order effects, and estimate heritabilities (proportion of observed variance due to genetic influences). Non-spatial alternation (NSA), which may involve greater reliance on processing of subgoals, was significantly more difficult than spatial alternation (SA). Despite their similarities, NSA and SA scores were uncorrelated. NSA performance was worse when administered second; there was no SA order effect. NSA scores were modestly heritable ($h^2=.25$; 26); SA was not. There was shared genetic variance between NSA scores and general intellectual ability ($r_g=.55$; .67), but this also suggests genetic influences specific to NSA. Compared with findings from small, selected control samples, high “failure” rates in this community-based sample raise concerns about interpretation of brain dysfunction in elderly or patient samples.

Keywords

Delayed Alternation; Spatial Alternation; Object Alternation; Non-Spatial Alternation; Working Memory; Set-Shifting; Prefrontal Cortex; Heritability; Twins

© 2011 Elsevier Inc. All rights reserved.

Correspondence: William S. Kremen, Ph.D., Department of Psychiatry, Center for Behavioral Genomics, University of California, San Diego, 9500 Gilman Drive (MC 0738), La Jolla, CA 92093-0738, Tel: 858-822-2393, Fax: 858-822-5856, wkremen@ucsd.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Portions of these data were presented by Tuan Mai at the annual meeting of the American Geriatrics Society, Seattle WA, May 2-6, 2007.

1. Introduction

The development of classic delayed response tasks to study working memory, mental set-shifting, and perseveration began nearly 100 years ago (Hunter, 1913; Jacobsen, 1936). Different versions have been studied in both animals and humans. These include non-human primates, healthy human volunteers, and people with brain disease or neuropsychiatric disorders (Curtis, Zald, Lee, & Pardo, 2000; Faraone et al., 2000; Freedman, Black, Ebert, & Binns, 1998; Freedman & Oscar-Berman, 1986; Fuster & Alexander, 1970; Gold, Berman, Randolph, Goldberg, & Weinberger, 1996; Goldman, Rosvold, Vest, & Galkin, 1971; Oscar-Berman, 1984; Oscar-Berman, Hutner, & Bonner, 1992; Oscar-Berman, McNamara, & Freedman, 1991; Oscar-Berman & Zola-Morgan, 1980; Oscar-Berman, Zola-Morgan, Oberg, & Bonner, 1982; Pribram, Mishkin, Rosvold, & Kaplan, 1952; Seidman et al., 1992; Turner & Levine, 2006; Zald, Curtis, Chernitsky, & Pardo, 2005; Zald, Curtis, Folley, & Pardo, 2002). Oscar-Berman (1980) referred to this paradigm as “comparative neuropsychology” because these measures constitute a set of relatively rare experimental tasks for study in humans that were developed as direct parallel to tasks tapping the same cognitive functions in non-human primates. These cognitive probes of prefrontal function are sensitive to brain lesions in monkeys and to homologous brain regions in humans (Freedman et al., 1998). They were developed, in part, to assess two major subsystems within the prefrontal cortex (Fuster, 1989; Oscar-Berman et al., 2009). The dorsolateral system is thought to mediate verbal and spatial working memory, set-shifting, attentional control, and concept formation. The orbitofrontal (ventral) system is thought to mediate response inhibition and emotional responsivity, and is more strongly linked to basal forebrain and limbic structures than the dorsolateral system. As such, these tests continue to be important in the study of frontal lobe function in normal and pathological conditions.

Differences still remain with regard to conclusions about test-specific or function-specific subregions including debate as to whether dorsolateral and ventral-inferior prefrontal cortex are distinguished primarily on the basis of spatial versus non-spatial delay, manipulation versus maintenance of information, response inhibition/resistance to interference, or reversal learning (Curtis & D’Esposito, 2004; Turner & Levine, 2006). Moreover, some basic properties of these tasks are still largely unknown. The purpose of the present study was to address some of these basic unresolved issues through the examination of two delayed response tasks that are variants of the widely used delayed alternation (DA) and object alternation (OA) tests. These issues include: 1) the extent to which genes influence performance; 2) what constitutes normal performance; 3) addressing the confound of interference suppression in comparing spatial and non-spatial alternation; 4) a comparison of level of difficulty between spatial and non-spatial alternation tasks; 5) how these tests are correlated with one another; and 6) whether order of test administration affects performance.

DA is a spatial alternation task and OA is a non-spatial alternation task. Figure 1 illustrates the variants of these tasks that are used in the present study. On each trial, participants have two choices to get a reward or a correct response. They must learn a “win-shift/lose-stay” strategy. In DA, the alternation (set-shifting) principle was based on spatial location such that a correct response on one side would result in the other side being correct on the next trial (hence, win-shift). OA—a non-spatial alternation test—was so named because in its original version, the alternation principle was based on “object” such that a correct response to one of two objects resulted in the other object being correct on the next trial. OA was somewhat more complicated than DA because the objects randomly switch from one location to the other on each trial. Therefore, in addition to the working memory and set-shifting components, OA had a cognitive interference component because any response to shifting location must be inhibited; only alternation between objects is relevant. We refer to the original, non-computerized versions of these tasks as classic DA and OA. As described

in the Methods section, our version of the non-spatial alternation test used color instead of object alternation.

1.1 Genetic influences on cognitive task performance

There is strong evidence for the importance of genetic factors in most cognitive functions (Bouchard & McGue, 2003). Little is known, however, about the heritability of many specific cognitive tests that are commonly used in neuropsychological or cognitive neuroscience research. The extent of genetic influences can vary dramatically across specific cognitive tests, and one cannot simply assume that because a particular cognitive domain has been found to be heritable, that any test within that domain must be heritable. Indeed, we have referred to this line of reasoning as the construct-measurement fallacy (Kremen & Lyons, 2010). For example, a review of IQ (*g*) measures indicates moderate-to-high heritability estimates ranging from about .50 to .70 (Bouchard & McGue, 2003). In contrast, in five of six studies of the Wisconsin Card Sorting Test (WCST) there was no or virtually no heritability across a wide variety of age ranges (Chou, Kuo, Lin, & Chen, 2009; Kremen, Eisen, Tsuang, & Lyons, 2007; Taylor, 2007). These results do not necessarily mean that cognitive functions tapped by the WCST are not heritable. For example, we found that reading span—a verbal working memory test—had a heritability of .51 (Kremen, Jacobsen et al., 2007). Also, Wright et al. (2002) reported a heritability of .45 for a simple delayed response spatial working memory test. It may, in part, be something about the psychometric properties of a test that determine whether or not it captures the genetic influences underlying particular cognitive functions. We are unaware of any studies of the heritability of the more complex delayed response tasks, but with the rapid growth in the field of genetics and genomics, the extent to which genetic factors influence performance on delayed response tasks is of crucial importance for a full understanding of working memory, set-shifting, and the prefrontal systems that underlie these functions.

1.2 Characterizing task performance in non-patients

As noted, DA and OA tasks in humans have been used previously to study brain disease/damage and neuropsychiatric disorders. These tasks are likely to be important for the study of normative cognitive and brain aging as well because working memory and executive functions are among the cognitive functions that are most sensitive to aging effects (Salthouse, 1991; Wingfield, Stine, Lahar, & Aberdeen, 1988), and cross-sectional structural MRI studies are consistent with the notion that the frontal lobes are the parenchymal brain region with the greatest age-related shrinkage (Raz, 2000). In order to fully understand these tasks and to understand normal as well as pathological changes in cognition and brain, it is necessary to examine these tasks in a large non-patient sample.

Sample sizes in most previous studies utilizing DA and OA tests in humans have been too small to make basic comparisons of specific features such as differences in task difficulty between the tests, the extent to which performance on the tests is correlated, or whether order of test administration affects performance.

1.3 Differences in spatial vs. object working memory, or in resistance to interference

The additional cognitive component requiring interference suppression in OA creates a confound that has not been addressed in previous animal or human studies using the classic DA and OA tests. In those studies, it was not possible to determine whether differences in DA and OA performance are attributable to differences in: 1) ability to process or shift between spatial (location) versus non-spatial (object) attributes (i.e., “where” vs. “what”); or 2) resistance to interference. In neuroimaging studies, Zald et al. (2005; 2002) and Turner and Levine (2006) used versions of DA that look identical to OA from the point of view of the study participant because objects randomly switch location from trial to trial in both

tests. In one, spatial location must be ignored and participants must alternate between objects. In the other, objects must be ignored and participant must alternate between spatial locations. In this way, the interference component is equivalent in both tests. The sole difference is in the alternation principle: spatial location (“where”) versus object (“what”). However, this difference between these tests and the classic DA tests was not addressed.

In the present study, we utilized similar tasks, but the alternation principle was based on color (as it was in the version developed by Gold et al. 1996). As shown in Figure 1, we refer to our variant of DA (which also includes an interference component) as spatial alternation (SA), and we refer to our variant of OA as non-spatial alternation (NSA). Our NSA task is much the same as the task used by Gold et al. Although they and others referred to their task as a hybrid delayed response/alternation test, we think it is best characterized as a variant of OA because it is a non-spatial alternation test.

1.4 Other issues

When DA and OA performance is compared, the focus has been on the cognitive processes that differentiate the tasks. Little attention has been paid to possible differences in task difficulty, but conclusions regarding differential deficits can be misleading if tasks are not equated for difficulty (Chapman & Chapman, 1978). Researchers and clinicians are often interested in how these tests are related to other cognitive tests, but we are unaware of any data on how they are actually related to one another. Finally, with alternation being their central focus, it is interesting that, to our knowledge, there are no data on order effects in the administration of DA and OA tasks. In the present study, we assessed our SA and NSA tasks in a large, non-patient twin sample in order to address the key issues that we have described.

2. Methods

2.1 Participants

Twins participated in a study of vulnerability to alcoholism in which they were randomly selected (with one exception) from 3322 twin pairs (6644 individuals) who had been interviewed by telephone in the Harvard Drug Study (Tsuang, Bar, Harley, & Lyons, 2001). The Harvard Drug Study enrolled all available participants from the Vietnam Era Twin Registry, a nationally distributed sample of male-male twin pairs in which both members served in the military at some time during the Vietnam era (1965–1975). The aforementioned exception was that only those without service in Vietnam were recruited for the alcohol vulnerability study. It should also be noted that the random selection meant that participants were not selected on the basis of substance use histories or any other demographic or clinical characteristics. The sample is similar to American men in their age range in terms of demographic and health characteristics (National Center for Disease Statistics, 2003).

Zygosity (monozygotic [MZ]; dizygotic [DZ]) was assigned to Registry members using questionnaire and blood group methods (Eisen, Neuman, Goldberg, Rice, & True, 1989) that have approximately 95% accuracy compared with DNA analysis (Nichols & Bilbro, 1966; Peeters, Van Gestel, Vlietinck, Derom, & Derom, 1998). Previous articles by Eisen et al. (1987) and Henderson et al. (1990) contain a complete description of the Registry’s construction.

To be included, both members of a pair had to agree to participate. There were 693 individuals in the present study; 690 were from pairs in which both members participated and 3 were unpaired twins whose co-twin initially agreed to participate but never actually did. Of those, there were 632 individuals (316 pairs) with valid SA and NSA scores for both

members of the pair. These participants were flown in from around the country for a day-long series of assessments at the University of California, Davis in Sacramento, CA and Harvard Medical School in Boston, MA. Participants were given their choice of study site. After complete description of the study to participants, written informed consent was obtained at the study sites.

There were 161 MZ and 155 DZ pairs. In virtually all cases, both members of a pair came together to the same site. The mean age of these participants was 47.9 years ($SD=3.3$; range=41–58); 92.2% were non-Hispanic white, 5.5% were African-American, 1.9% were Hispanic, and 0.4% were other; 97% graduated high school or obtained a GED, 33% were college graduates; 98% were employed full-time and 1.7% were employed part-time.

2.2 Delayed response tests

The SA and NSA tests in the present study had the same timing, rule acquisition principles, and criterion for passing or failing as the DA and OA tasks created by Oscar-Berman and Zola-Morgan (Oscar-Berman & Zola-Morgan, 1980). Unlike the original tests, ours were computer-administered, the stimuli were different, and there was an interference component to the SA task. The tasks were programmed in C++ by Howard Blankfeld.

Instead of the classic wells, a red and a blue square appeared side by side on the monitor in both tests. Participants were asked to select one of the squares by pressing the left mouse button for the left square or the right mouse button for the right square. Testers pointed to the relevant buttons and squares while giving the instructions because we were concerned that use of the words “left” and “right” might function as an unintended cue to attend to spatial location. Participants were then told that after each selection, the word “Correct” or “Incorrect” would appear on the screen. It was explained that no further information could be provided, but their goal was to use this feedback in order to get as many right and as few wrong as possible. Finally, they were instructed to work both as quickly and as accurately as possible. On the second test given, participants were told that “this one is a little bit different” but that the instructions were the same, i.e., to use the feedback to get as many right and as few wrong as possible.

The first response was always correct. There was a 5-second delay interval between trials. The test was discontinued when a participant achieved the learning criterion of 12 correct consecutive responses or the failure criterion of 50 total incorrect responses anywhere on the test. The test order was counterbalanced across twin pairs but the order of administration was always the same within a twin pair. In SA, the to-be-learned alternation pattern was that after each correct response, the spatial location of the next correct response shifted. In NSA, the to-be-learned alternation pattern was that after each correct response, the color of the next correct response shifted. Examples of a series of SA and NSA trials are shown in Figure 1. Scores used for analyses were the total number of errors, and a dichotomous score based on whether or not a participant achieved the learning criterion.

2.3 Statistical analysis

Because the distributions of the total error scores deviated substantially from normal, these scores were converted into ordinal levels so that polychoric correlations could be calculated. Polychoric correlations assume underlying normal distribution. The contingency tables can be tested for their fit to a bivariate normal distribution in Mx, a maximum likelihood-based structural equation modeling package that is specialized for twin analyses (Neale, Boker, Xie, & Maes, 2003). We used 5 levels for SA errors and 4 levels for NSA errors; both contingency tables did provide an acceptable fit to the bivariate normal distribution. Comparisons of dichotomous variables were tested with χ^2 statistics. The number of

participants reaching the learning criterion on each test was compared by means of a test of difference in proportions (McNemar, 1975).

Twin analyses—MZ twins share 100% of their genes, and DZ twins on average share 50% of their genes. This relationship makes it possible to decompose phenotypic variation in the basic twin model into three latent sources of variance: additive genetic influences (A); shared or common environmental influences (C); and nonshared or individual-specific environmental influences (E) (Eaves, Last, Young, & Martin, 1978; Neale & Cardon, 1992). Relationships across twins are modeled as follows: 1) Based on the differences in gene sharing, additive genetic factors correlate 1.0 for MZ twins and 0.5 for DZ twins. 2) Shared environmental influences (factors that contribute to twin similarity) correlate 1.0 within twin pairs regardless of zygosity. 3) Nonshared environmental influences (factors that contribute to differences between twins) are uncorrelated within twin pairs. Measurement error is included in the nonshared environmental variance; given that error is assumed to be random, it is also assumed to be uncorrelated between twins.

A saturated model that perfectly recaptures the observed means, variances, and covariances within each zygosity group exactly was fit to the raw data. The saturated model fit was then used as a comparison model to assess the overall goodness-of-fit of each of the models of interest. It was first compared to the full model; this is the classic “ACE” model reflecting all three components of variance. Reduced models test whether dropping A, C, or both components results in a significant worsening of fit. Because there is always measurement error, it is not possible to have a model without an E component. If MZ correlations are substantially more than double the DZ correlations, it suggests non-additive genetic variance. In this case, the more appropriate model is an “ADE” model in which D reflects non-additive/dominance genetic variance. C and D cannot be included in the same model because it would be over-identified.

Model comparisons were based on the likelihood-ratio chi-square test (LRT) statistic and the Akaike Information Criterion (AIC; Akaike, 1987). The LRT is obtained by calculating the difference between the $-2 \log$ -likelihood ($-2LL$) of a comparison model and the $-2LL$ of a nested (reduced) model. A nonsignificant LRT is generally accepted as indicating that the reduced model is the better (i.e., more parsimonious) model because it is able to account for the data with fewer parameters. When the LRT is nonsignificant for two or more competing models, the AIC is used to determine the preferred model. The AIC, calculated as $\Delta\chi^2 - 2\Delta df$, indexes both goodness-of-fit and parsimony. The more negative the AIC, the better the balance between goodness-of-fit and parsimony. Squaring the parameter estimates of the A, C, D, and E influences on the phenotype yields the proportion of variance accounted for by additive genetic (a^2), shared environmental (c^2), non-additive genetic (d^2), and nonshared environmental (e^2) influences, respectively. In the ACE model, the a^2 term is also what is commonly referred to as the heritability. Readers who may be more familiar with term h^2 should note that a^2 and h^2 are interchangeable in this instance. Heritability in the ADE model is based on the total genetic variance ($a^2 + d^2$). Our sample is substantially underpowered to differentiate between A and D effects (Martin, Eaves, Kearsley, & Davies, 1978; Neale, Eaves, & Kendler, 1994; Visscher, 2006). Therefore, it is most appropriate to ignore differences in a^2 and d^2 and to focus on the broad sense heritability ($a^2 + d^2$).

These univariate models are easily extended to the bivariate case. We can then not only examine these sources of variance, but also the sources of covariance between measures. Based on our results (see below), we tested bivariate genetic models between one of our delayed response measures and an estimate of general intellectual ability. For purposes of the present study, the primary focus of these analyses was to compute genetic correlations between these measures. A phenotypic correlation is based on the covariance between two

measures; it includes the combined genetic and environmental covariance. A genetic correlation is based only on the genetic covariance. It can be computed because it is possible to decompose genetic and environmental sources of variance in the twin design. It is defined as the genetic covariance divided by the square root of the product of the genetic variance of one variable and the genetic variance of the other variable. Note that removing the word “genetic” leaves one with the definition of a standard, phenotypic correlation. Unique environmental correlations are analogous to the genetic correlations with respect to unique environmental variances.

3. Results

3.1 Non-twin analyses

A significantly higher proportion of participants achieved the learning criterion on SA compared with NSA (79% vs. 50%, $z = 6.75$, $p < .05$). The mean number of errors was 20.68 (SD=19.03) on SA and 31.27 (SD=20.51) on NSA. SA performance was unrelated to NSA performance, whether it was examined on the basis of achieving the learning criterion ($\chi^2[1]=0.45$, tetrachoric $r = -.05$, $p = .50$) or total errors (polychoric $r = .06$, $p = .29$). The proportion of participants achieving the learning criterion indicated that there was no order effect for SA: 80% of those who were administered SA first versus 78% of those who were administered SA second ($\chi^2[1]=0.42$, $p = .52$). In contrast, there was a significant order effect for NSA: 56% of those who were administered NSA first versus 45% of those were administered NSA second ($\chi^2[1]=7.28$, $p < .007$). There was no time-of-day effect (morning vs. afternoon) for either test. Both $\chi^2[1]$ values were < 1 ($ps > .41$).

In our sample, the reading subtest of the Wide Range Achievement Test – Version 3 (WRAT-3; Wilkinson, 1993) was used as an estimate of general intellectual ability (cf. Kremen et al., 1996). Phenotypic Spearman rank-order correlations with reading were $-.21$ ($p < .0001$) for NSA errors and $-.08$ ($p < .04$) for SA errors. Absolute values of Spearman correlations with NSA errors ranged from .09 to .18 ($.03 < ps < .001$) for measures of attention, visual-spatial function, verbal and visual-spatial declarative memory, and other executive functions. Better scores on one test were always associated with better scores on the other. Correlations were similar but slightly smaller for SA. Although there may be age effects on the SA and NSA tests, they were not manifest in the midlife age range of our sample ($rs < .07$). Nor were there significant correlations ($rs < .07$) with self-reported general physical and mental health on the Medical Outcome Study Short Form-36 (Ware & Sherbourne, 1992).

3.2 Twin analyses

The MZ and DZ correlations were generally low for both tests (shown in Table 1 along 95% confidence intervals [CIs]). As would be expected, the patterns were very similar for achieving the criterion and total errors. Neither SA score was heritable because in both cases the MZ correlations were actually somewhat (although not significantly) lower than the DZ correlations. Table 2 shows the model comparison results for SA. It indicates that the most parsimonious (best-fitting model) was an E only model (5a) for achieving the criterion and a CE model (4b) for errors. The E only model is consistent with the fact that DZ correlations were somewhat higher than MZ correlations for achieving the SA criterion. That is, no theoretical model would predict this pattern, so the best interpretation is that it reflects nonshared environmental influences and measurement error. In the CE model for SA errors, C accounted for 11% (.00, .23) of the variance and E accounted 89% (.77, 1.00) of the variance (see Table 4). The 95% CI for the C estimate was bounded at zero, dropping C from the model still resulted in a model with a good fit; however, the CE model was the most parsimonious model based on the AIC.

Table 3 shows the model comparisons for NSA. Based on the MZ and DZ correlations, we included an ADE model. The AIC values indicate that the ADE model for both achieving the criterion and total errors had a better fit to the data than the ACE model. However, the AE models (4a and 4b) were the most parsimonious model for both scores. Heritability estimates (see Table 4) in these best-fitting models were similar for both measures: 25% (.03, .45) for achieving the criterion; and 26% (.09, .43) for total errors. All parameters for the full and most parsimonious models for the SA and NSA variables are shown in Table 4.

Given the phenotypic correlation between estimated general intellectual ability (WRAT-3 reading) and NSA performance, we performed bivariate genetic analyses to determine the extent to which genetic or environmental covariance underlies this association. Because the NSA variables were dichotomous or ordinal, we created an ordinal WRAT-3 reading measure with 8 levels. We then carried out the bivariate correlated factors analyses with A, C, and E parameters for WRAT reading, but only A and E parameters for the NSA measures (because the C estimates were zero). Signs were reversed for NSA errors so that all positive correlations indicate better NSA performance being associated with higher estimated intellectual ability. Results for NSA errors are shown in Figure 2. Heritability of the ordinal WRAT-3 reading measure was .46 (95% CI: .23; .71), consistent with the previously published estimate of .45 based on the continuous measure (Kremen et al., 2005). There was a good overall fit of the model (-2LL=3417.784; df=1325; LRT=0.045; Δ df=2, p=.978, AIC=-3.955). There was a significant genetic correlation of .67 (95% CI: .32; 1.00), and a nonsignificant unique environmental correlation of .05 (95% CI: -.15; .24). Results were very similar for NSA errors. The model fit statistics for NSA errors were as follows: -2LL=4096.787; df=1321; LRT=0.084; Δ df=2, p=.959, AIC=-3.916. There was a significant genetic correlation of .55 (95% CI: .24; 1.00), and a nonsignificant unique environmental correlation of .11 (95% CI: -.06; .29).

There was a good overall fit of the model (-2LL=4096.787; df=1321; LRT=0.084; Δ df=2, p=.959, AIC=-3.916). There was a significant genetic correlation of .55 (95% CI: .24; 1.00), and a nonsignificant unique environmental correlation of .12 (95% CI: -.06; .29). Results were very similar for achieving the learning criterion. The model fit statistics for achieving the learning criterion were as follows: -2LL=3417.784; df=1325; LRT=0.045; Δ df=2, p=.978, AIC=-3.955. There was a significant genetic correlation of .67 (95% CI: .32; 1.00), and a nonsignificant unique environmental correlation of .05 (95% CI: -.16; .24).

As a check on specific genetic variance, we examined the genetic variance that was specific to NSA errors, after accounting for the genetic influences on WRAT-3. The specific genetic variance was .19. The genetic variance that was specific to WRAT-3, after accounting for genetic influences in NSA errors, was .31. These results indicate that nearly 70% of the genetic variance for each measure (.19 of .27; and .31 of .46) was accounted for by genetic influences specific to that measure. However, neither was significant based on the 95% confidence intervals; indeed, they could not be significant with the upper bound of the confidence interval for the genetic correlation being 1.00. The same pattern was present for achieving the learning criterion in that the specific genetic variances each accounted for over one-half of the total genetic variance (.14 of .25; and .24 of .45), but neither was significant.

4. Discussion

4.1 Behavior genetics

The behavior genetic analyses indicated modest heritability for NSA, but no heritability for SA. As noted in the introduction, there is evidence for the heritability of executive and working memory functions, but that is not the case for all tests measuring these abilities. It appears rather difficult to predict in advance which measures will or will not be significantly

heritable, and we do not know what would make NSA, but not SA, heritable. Although the heritability estimates for SA measures were .00, it should also be noted that the 95% confidence intervals for SA went as high as .25. Our sample is not large for a twin study; hence, we cannot conclude with a high degree of certainty that there is no heritable variation in SA in the general population. On the other hand, we previously reported a heritability estimate of .00 for the WCST (Kremen, Eisen et al., 2007), and as noted in the Introduction, almost all other WCST studies have been consistent with our finding. Thus, it is not unreasonable to expect a similar heritability estimate in other studies of SA. Given the upper limit of the confidence interval, it is unlikely to exceed .25, but only future replications can provide confirmation. Because cognitive measures that are not heritable are not likely to be useful for studies that are seeking to differentiate performance as a function of different genotypes, one important implication of the results of these twin analyses is that SA performance may not be a useful cognitive phenotype for genetic association studies. At least based on these initial findings, it would be unwise to select SA as a phenotype in a genetic association study.

Our bivariate genetic analysis revealed a genetic correlation of .67 between NSA and estimated general intellectual ability based on WRAT-3 reading performance. This association indicates a fair amount of shared genetic variance between the two, and is consistent with the well-known association between general intellectual ability and working memory/executive function. On the other hand, being far less than 1.00, this genetic correlation also indicates that there are likely to be some genetic influences that are specific to NSA performance and independent of general intellectual ability. This interpretation is complicated by the fact that genetic influences that were specific to NSA performance were not significant after adjusting for the genetic influences on our reading-based estimate of general intellectual ability. It might be tempting to conclude that the lack of significant specific genetic influences on NSA means that it really represents primarily genetic influences on general intellectual/reading ability, but the heritability of our general intellectual ability/reading measure was also no longer significant after adjusting for genetic influences on NSA performance. Thus, one would also have to conclude that our general intellectual ability/reading measure really represents primarily working memory and executive function. The latter seems unlikely. Furthermore, it seems unlikely that there are actually no specific genetic influences on NSA when 56% of the genetic variance in achieving the learning criterion and 70% of the genetic variance in errors were specific to NSA. Therefore, it seems most likely that there are true specific effects but the confidence intervals indicate that we are underpowered to detect them.

4.2 Spatial and non-spatial alternation: What is abnormal?

We examined SA and NSA performance in a large middle-aged male twin sample. In non-twin analyses, 79% of the sample achieved the learning criterion on SA but only 50% achieved the criterion on NSA. These results are important with regard to the definition of normal or abnormal when it comes to such working memory, rule-learning tasks. When these tests are used in neuropsychological studies comparing patients and controls, the assumption tends to be that virtually all relatively healthy individuals (i.e., people without brain disease or damage, or serious neuropsychiatric disorders) should perform successfully on these tests.

In our results in a large sample of community-dwelling middle-aged men, fully one-half of the participants were unable to achieve the learning criterion on the NSA test. If one-half of this community-based, middle-aged sample was unable to successfully perform this test, then it may be exceedingly difficult to make inferences about compromised cognitive or brain function when failures are observed in neurologic or neuropsychiatric patients. Such inferences may be even more difficult in patients with age-related diseases (e.g., Parkinson's

disease) because it is not unlikely that the percentage of people who fail to reach the learning criterion would be even higher in a community sample of elderly adults in comparison to the present middle-aged sample. Only about a fifth of our sample failed to achieve the learning criterion on SA, but that frequency is still high enough to raise concerns about false positive identification of impairment in patient samples.

Previous studies provide a mixed picture regarding consistency with our results, but comparisons are limited because most had very small, screened samples. First, we consider performance when participants were given the rules in advance. Two neuroimaging studies reported mean OA accuracy rates of over 95% (Turner & Levine, 2006; Zald et al., 2002), but a study using a color alternation task reported accuracy of only 69% (Gold et al., 1996). In the latter study, mean accuracy rose to 99% during a second testing. Thus, it is unclear whether the difference in results may be a function of task novelty or color versus object working memory. Next, we consider studies in which the rule had to be figured out. In a neuroimaging study, only 6 of 10 participants achieved the learning criterion (Zald et al., 2005).

A rare study with a reasonably large number of controls examined the classic OA task in three control groups that totaled up to 70 individuals; a weighted average of these groups resulted in a mean age of 56.1 years, mean IQ of 115.4, and a mean of 24.3 errors (Dirksen, Howard, Cronin-Golomb, & Oscar-Berman, 2006). The mean of 24.3 errors is somewhat lower than our mean of 31.3, but general intellectual ability may be a key difference. In our sample, the mean estimated general intellectual ability based on WRAT-3 reading was at the average level (97.21; SD=10.96) compared to the above average IQ of 115 in the Dirksen et al. sample. This comparison, together with our bivariate genetic results, supports the notion that the performance difference between our sample and the Dirksen et al. sample is likely to be due largely to differences in general intellectual ability. Thus, although the level of difficulty experienced by participants in the present study might initially seem surprising, these comparisons suggest that it is not anomalous.

4.3 Comparing SA and NSA performance

Because the stimuli and motor responses were identical on each test, it is very surprising that SA and NSA performances were entirely uncorrelated. The low variability in SA performance (e.g., 79% achieving the learning criterion) could be a factor in there being low correlations, but this pattern warrants further study. Despite their similarities, NSA was also significantly more difficult than SA. In studies using classic DA and OA, greater difficulty of OA might be attributed to the non-spatial alternation component (vs. the spatial alternation of DA) or to the interference component that is intrinsic to OA. However, it would not be possible to choose one explanation over the other because the tests differed on both dimensions. In contrast, the SA test in the present study did have an interference component that was analogous to that of NSA. It appears, therefore, that the difference in these two tests must be a function of the spatial versus non-spatial alternation rather than interference suppression component. We consider our results to be consistent with those of Turner and Levine (2006) where accuracy was higher on DA than on OA ($d=.63$). Although this comparison was not statistically significant with only 10 participants, the effect size does suggest a meaningful difference. On the other hand, Zald et al. (2002), who also utilized an SA task, had ceiling effects on both tests. Our results highlight the fact that carefully equating all components of different cognitive tests does not necessarily translate into equivalent difficulty.

Having included an interference component in both tests, the question of why non-spatial alternation would be more difficult than spatial alternation remains unanswered. One possibility is that the difference hinges on the phenomenon of stimulus-response

congruence. There is always an automatic stimulus-response congruence on all trials in SA; for example, if the correct stimulus is on the left, the correct response is always the left button. Indeed, once a person figures out the alternation principle on SA, the test could be done while barely paying any attention to the stimuli at all. One would know that button presses simply needed to be alternated on every trial. In contrast, NSA trials do not possess an automatic stimulus-response congruence; if the correct stimulus is red, the correct response may be either left or right on any given trial. One way to summarize this distinction is that an additional subgoal (i.e., more top-down control) is required in NSA. It should also be noted that our version of NSA is actually a color alternation test like the task developed by Gold et al. (1996) in contrast to the object alternation tasks developed by Zald et al. (2002) and Turner and Levine (2006). The classic versions that were developed by Oscar-Berman and colleagues (1980) consisted of objects that differed in both color and shape. As such, one cannot know whether participants in studies using that apparatus were shifting according to color or shape. With color, shape, and interference effects being equivalent in both of our tasks, we have assumed that they provide a comparison of spatial versus non-spatial working memory.

We also examined difficulty with respect to order effects. Here we refer to test order rather than trial-by-trial switch effects. There were no apparent fatigue effects as performance levels did not differ as a function of testing earlier or later in the day. There was no evidence for practice effects because, when there were differences, performance on the test taken second was worse than when it was taken first. Learning the alternation principle on the first test means that it would have become strongly primed. Participants would then need to inhibit the now prepotent response tendency in order to succeed on the subsequent test. If this Stroop-like phenomenon was at work, it seems to have occurred only when going from SA to NSA. Although we cannot definitively account for the order effect, our results do indicate that the impact of order effects should be taken into account when studying these types of tasks.

4.5 Strengths and limitations

A limitation of this study was that the sample was made up of mostly Caucasian males, so results may not generalize to women and ethnic minorities. Although the study participants were similar to American men in their age range, it would still be useful to examine these tests in non-veterans. Also, it is possible that performance on the delayed response tasks was affected by verbal mediation. There is evidence of a female advantage in some verbal abilities, but we are unaware of much evidence for sex differences in executive functions; however, it does seem that any verbal mediation on these tests does not require a high level of verbal ability. In addition, it is possible that the results are largely accounted for by error variance given that error is included in the unique environment term which accounted for most of the variance. However, this possibility seems rather unlikely considering the meaningful patterns of results in the literature using very similar tests. There is also a risk of circular reasoning if one assumes that any phenotype with low heritability must have substantial error variance. Given the fact that 79% of participants achieved the learning criterion on SA and that we found significant heritability for NSA, it seems unlikely that scores are due largely to random (i.e., error) factors.

The large sample size is a major strength of the study that permitted comparisons of difficulty level, associations between tests, order effects, and behavior genetic analyses; however, we were still underpowered for some genetic analyses. It is also informative to observe performance on these tests in a community-dwelling, non-patient sample that was not highly screened. The results suggest that acquiring a rule of non-spatial alternation is more difficult than acquiring a rule of spatial alternation. The results also raised questions about definitions of abnormal performance because only one-half of the participants in this

non-patient sample achieved the learning criterion in the NSA task. Also, because SA was not heritable, the results suggest that NSA, but not SA, may be useful genetic association studies. Bivariate analyses suggested that some genes that influence general intellectual ability are also likely to influence NSA performance; however, the results also suggested that there are genetic influences that are specific to NSA performance.

The results are also relevant for the examination of frontal-executive functions in patient groups and in studies of normal and pathological aging. There is ample evidence that working memory and executive functions are among the more age-susceptible cognitive functions, and that there is greater age-associated frontal lobe shrinkage relative to other brain regions (Hedden & Gabrieli, 2004). Thus, although the relatively narrow age range of the present study may preclude generalization to other age groups, it can nevertheless be stated with some confidence that older adults are likely to perform more poorly on these tests. Heritability can also change with age, and it is, therefore, possible that the heritability of performance on these tests may be different in different age groups.

Acknowledgments

Preparation of this article was supported by National Institute of Alcohol Abuse and Alcoholism AA10586 and National Institute on Aging Grants AG18384, AG18386, AG22381, and AG22982, and by the National Institute on Aging, American Federation for Aging Research, and the Hartford Foundation through the UCSD Stein Institute for Research on Aging.

The U.S. Department of Veterans Affairs has provided financial support for the development and maintenance of the Vietnam Era Twin (VET) Registry. Numerous organizations have provided invaluable assistance in the conduct of this study, including: Department of Defense; National Personnel Records Center, National Archives and Records Administration; Internal Revenue Service; National Opinion Research Center; National Research Council, National Academy of Sciences; the Institute for Survey Research, Temple University. Most importantly, the authors gratefully acknowledge the continued cooperation and participation of the members of the VET Registry and their families. Without their contribution this research would not have been possible.

References

- Akaike H. Factor analysis and AIC. *Psychometrika*. 1987; 52:317–332.
- Bouchard TJ Jr, McGue M. Genetic and environmental influences on human psychological differences. *Journal of Neurobiology*. 2003; 54:4–45. [PubMed: 12486697]
- Chapman LJ, Chapman JP. The measurement of differential deficit. *Journal of Psychiatric Research*. 1978; 14:303–311. [PubMed: 722633]
- Chou LN, Kuo PH, Lin CC, Chen WJ. Genetic and environmental influences on the Wisconsin Card Sorting Test performance in healthy adolescents: A twin/sibling study. *Behavior Genetics*. 2009
- Curtis CE, D'Esposito M. The effects of prefrontal lesions on working memory performance and theory. *Cognitive, Affective and Behavioral Neuroscience*. 2004; 4:528–539.
- Curtis CE, Zald DH, Lee JT, Pardo JV. Object and spatial alternation tasks with minimal delays activate the right anterior hippocampus proper in humans. *Neuroreport*. 2000; 11:2203–2207. [PubMed: 10923671]
- Dirksen CL, Howard JA, Cronin-Golomb A, Oscar-Berman M. Patterns of prefrontal dysfunction in alcoholics with and without Korsakoff's syndrome, patients with Parkinson's disease, and patients with rupture and repair of the anterior communicating artery. *Neuropsychiatr Disease and Treatment*. 2006; 2:327–339.
- Eaves LJ, Last KA, Young PA, Martin NG. Model-fitting approaches to the analysis of human behavior. *Heredity*. 1978; 41:249–320. [PubMed: 370072]
- Eisen SA, Neuman R, Goldberg J, Rice J, True W. Determining zygosity in the Vietnam Era Twin Registry: An approach using questionnaires. *Clinical Genetics*. 1989; 35:423–432. [PubMed: 2736790]

- Eisen SA, True WR, Goldberg J, Henderson W, Robinette CD. The Vietnam Era Twin (VET) Registry: Method of construction. *Acta Geneticae Medicae et Gemellologiae*. 1987; 36:61–66. [PubMed: 3673478]
- Faraone SV, Seidman LJ, Kremen WS, Toomey R, Pepple JR, Tsuang MT. Neuropsychological functioning among the nonpsychotic relatives of schizophrenic patients: The effect of genetic loading. *Biological Psychiatry*. 2000; 48:120–126. [PubMed: 10903408]
- Freedman M, Black S, Ebert P, Binns M. Orbitofrontal function, object alternation and perseveration. *Cerebral Cortex*. 1998; 8:18–27. [PubMed: 9510382]
- Freedman M, Oscar-Berman M. Comparative neuropsychology of cortical and subcortical dementia. *Canadian Journal of Neurological Science*. 1986; 13:410–414.
- Fuster JM. The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe. 2. Vol. 2. New York: Raven Press; 1989.
- Fuster JM, Alexander GE. Delayed response deficit by cryogenic depression of frontal cortex. *Brain Research*. 1970; 20:85–90. [PubMed: 4986430]
- Gold JM, Berman KF, Randolph C, Goldberg TE, Weinberger DR. PET validation of a novel prefrontal task: Delayed response alternation. *Neuropsychology*. 1996; 10:3–10.
- Goldman PS, Rosvold HE, Vest B, Galkin TW. Analysis of delayed alternation deficit produced by dorsolateral prefrontal lesions in the rhesus monkey. *Journal of Comparative and Physiological Psychology*. 1971; 77:212–220. [PubMed: 5000659]
- Hedden T, Gabrieli JD. Insights into the ageing mind: A view from cognitive neuroscience. *Nature Reviews Neuroscience*. 2004; 5:87–96.
- Henderson WG, Eisen SE, Goldberg J, True WR, Barnes JE, Vitek M. The Vietnam Era Twin Registry: A resource for medical research. *Public Health Reports*. 1990; 105:368–373. [PubMed: 2116638]
- Hunter WS. The delayed reaction in animals and children. *Behavior Monograph*. 1913; 2:1–13.
- Jacobsen CF. Studies of cerebral function in primates. *Comparative Psychology Monographs*. 1936; 13:1–68.
- Kremen WS, Eisen SA, Tsuang MT, Lyons MJ. Is the Wisconsin Card Sorting Test a useful neurocognitive endophenotype? *American Journal Medical Genetics (Neuropsychiatric Genetics)*. 2007; 144B:403–406.
- Kremen WS, Jacobsen KC, Xian H, Eisen SA, Eaves LJ, Tsuang MT, et al. Genetics of verbal working memory processes: A twin study of middle-aged men. *Neuropsychology*. 2007; 21:569–580. [PubMed: 17784805]
- Kremen WS, Jacobson KC, Xian H, Eisen SA, Waterman B, Toomey R, et al. Heritability of word recognition in middle-aged men varies as a function of parental education. *Behavior Genetics*. 2005; 35:417–433. [PubMed: 15971023]
- Kremen WS, Lyons MJ. Behavior genetics of aging. In: Schaie, KW.; Willis, SL., editors. *Handbook of the psychology of aging*. 7. San Diego, CA: Elsevier; 2010. p. 93–107.
- Kremen WS, Seidman LJ, Faraone SV, Pepple JR, Lyons MJ, Tsuang MT. The “3 Rs” and neuropsychological function in schizophrenia: An empirical test of the matching fallacy. *Neuropsychology*. 1996; 10:22–31.
- Martin NG, Eaves LJ, Kearsy MJ, Davies P. The power of the classical twin study. *Heredity*. 1978; 40:97–116. [PubMed: 272366]
- McNemar, Q. *Psychological Statistics*. New York: Wiley; 1975.
- National Center for Disease Statistics. Health, United States, 2003. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2003.
- Neale, MC.; Boker, SM.; Xie, G.; Maes, HH. *Mx: Statistical Modeling*. 6. Richmond, VA: Department of Psychiatry, Medical College of Virginia; 2003.
- Neale, MC.; Cardon, LR. *Methodology for genetic studies of twins and families*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1992.
- Neale MC, Eaves LJ, Kendler KS. The power of the classical twin study to resolve variation in threshold traits. *Behavior Genetics*. 1994; 24:239–258. [PubMed: 7945154]

- Nichols RC, Bilbro WCJ. The diagnosis of twin zygosity. *Acta Genetica et Statistica Medica*. 1966; 16:265–275. [PubMed: 4959152]
- Oscar-Berman, M. Comparative neuropsychology and alcoholic Korsakoff disease. In: Squire, L.; Butters, N., editors. *Neuropsychology of memory*. New York: Guilford; 1984. p. 194–202.
- Oscar-Berman M, Hutner N, Bonner RT. Visual and auditory spatial and nonspatial delayed-response performance by Korsakoff and non-Korsakoff alcoholic and aging individuals. *Behavioral Neuroscience*. 1992; 106:1–10.
- Oscar-Berman, M.; McNamara, P.; Freedman, M. Delayed response tasks: Parallels between experimental ablation studies and findings in patients with frontal lesions. In: Levin, HS.; Eisenberg, HM.; Benton, AL., editors. *Frontal lobe function and dysfunction*. New York: Oxford University Press; 1991. p. 230–255.
- Oscar-Berman M, Valmas MM, Sawyer KS, Kirkley SM, Gansler DA, Merritt D, et al. Frontal brain dysfunction in alcoholism with and without antisocial personality disorder. *Neuropsychiatric Disease and Treatment*. 2009; 5:309–326. [PubMed: 19557141]
- Oscar-Berman M, Zola-Morgan SM. Comparative neuropsychology and Korsakoff's syndrome. I. Spatial and visual reversal learning. *Neuropsychologia*. 1980; 18:499–512. [PubMed: 6449680]
- Oscar-Berman M, Zola-Morgan SM, Oberg RGE, Bonner RT. Comparative neuropsychology and Korsakoff's syndrome. III: Delayed response, delayed alternation and DRL performance. *Neuropsychologia*. 1982; 20:187–202. [PubMed: 6211634]
- Peeters H, Van Gestel S, Vlietinck R, Derom C, Derom R. Validation of a telephone zygosity questionnaire in twins of known zygosity. *Behavior Genetics*. 1998; 28:159–163. [PubMed: 9670591]
- Pribram KH, Mishkin M, Rosvold HE, Kaplan S. Effects of delayed response performance of lesions of dorsolateral and ventromedial frontal cortex of baboons. *Journal of Comparative and Physiological Psychology*. 1952; 46:565–572. [PubMed: 13000029]
- Raz, N. Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In: Craik, FIM.; Salthouse, TA., editors. *The Handbook of Aging and Cognition*. Hillsdale, NJ: Erlbaum; 2000. p. 1–90.
- Salthouse TA. Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science*. 1991; 2:179–183.
- Seidman LJ, Talbot NL, Kalinowski AG, McCarley RW, Faraone SV, Kremen WS, et al. Neuropsychological probes of frontal-limbic dysfunction in schizophrenia: Olfactory identification and Wisconsin Card Sorting performance. *Schizophrenia Research*. 1992; 6:55–65. [PubMed: 1786234]
- Taylor J. Heritability of Wisconsin Card Sorting Test (WCST) and Stroop Color-Word Test performance in normal individuals: Implications for the search for endophenotypes. *Twin Research and Human Genetics*. 2007; 10:829–834. [PubMed: 18179394]
- Tsuang MT, Bar JL, Harley RM, Lyons MJ. The Harvard Twin Study of Substance Abuse: What we have learned. *Harvard Review of Psychiatry*. 2001; 9:267–279. [PubMed: 11600486]
- Turner GR, Levine B. The functional neuroanatomy of classic delayed response tasks in humans and the limitations of cross-method convergence in prefrontal function. *Neuroscience*. 2006; 139:327–337. [PubMed: 16324791]
- Visscher PM. A note on the asymptotic distribution of likelihood ratio tests to test variance components. *Twin Research and Human Genetics*. 2006; 9:490–495. [PubMed: 16899155]
- Ware JE, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care*. 1992; 30:473–483. [PubMed: 1593914]
- Wilkinson, GS. WRAT 3: Wide Range Achievement Test Manual. Wilmington, DE: Wide Range, Inc; 1993.
- Wingfield A, Stine EAL, Lahar CJ, Aberdeen JS. Does the capacity of working memory change with age? *Experimental Aging Research*. 1988; 14:103–107. [PubMed: 3234452]
- Wright MJ, Luciano M, Hansell NK, Geffen GM, Geffen LB, Martin NG. Genetic sources of covariation among P3(00) and online performance variables in a delayed-response working memory task. *Biological Psychology*. 2002; 61:183–202. [PubMed: 12385675]

- Zald DH, Curtis C, Chernitsky LA, Pardo JV. Frontal lobe activation during object alternation acquisition. *Neuropsychology*. 2005; 19:97–105. [PubMed: 15656767]
- Zald DH, Curtis C, Folley BS, Pardo JV. Prefrontal contributions to delayed spatial and object alternation: A positron emission tomography study. *Neuropsychology*. 2002; 16(2):182–189. [PubMed: 11949710]

Figure 1. Sample Trial Sequences for Spatial Alternation (SA) and Non-Spatial Alternation (NSA) Tests. SA is analogous to classic delayed alternation tests and NSA is analogous to classic object alternation tests.

Figure 2.

Bivariate Correlated Factors Model for Estimated General Intellectual Ability (WRAT-3 Reading) and NSA Performance. WRAT-3 = Wide Range Achievement Tests – Version 3; NSA = Non-Spatial Alternation Test; r_g = Genetic correlation; r_e = Unique environmental correlation; A = Additive genetic influences; C = Common environmental influences; E = Unique environmental influences. Estimate shown are squared parameter estimates (indicating percent variance accounted for) and 95% confidence intervals. Signs were reversed for NSA errors so that all positive correlations indicate better NSA performance being associated with higher estimated intellectual ability.

Table 1

MZ and DZ correlations for Spatial Alternation (SA) and Non-Spatial Alternation (NSA) tests

	r_{MZ} (95% confidence interval)	r_{DZ} (95% confidence interval)
SA: Achieve Criterion ^a	-.16 (-.44, .14)	.35 (.05, .64)
NSA: Achieve Criterion ^a	.29 (.05, .50)	.03 (.00, .27)
SA: Total Errors ^b	.05 (.0006, .22)	.19 (.02, .35)
NSA: Total Errors ^b	.29 (.10-.46)	.08 (.00-.23)

MZ=Monozygotic; DZ=Dizygotic.

^aTetrachoric correlations;^bPolychoric correlations.

Table 2

Model comparisons for Spatial Alternation (SA) test

Model	-2LL	df	LRT	Δdf	p	AIC
<u>SA Achieving Criterion</u>						
1a Saturated Model	664.796	650	-	-	-	-
2a ACE Model	671.623	653	6.827	3	.078	0.827
3a AE Model	672.292	654	7.496	4	.112	-0.504
4a CE Model	671.623	654	6.827	4	.145	-1.173
5a E Model	672.292	655	7.496	5	.186	-2.504
<u>SA Total Errors</u>						
1b Saturated Model	2084.208	636	-	-	-	-
2b ACE Model	2099.900	648	15.692	12	.206	-8.308
3b AE Model	2101.149	649	16.941	13	.202	-9.059
4b CE Model	2099.899	649	15.691	13	.266	-10.309
6b E Model	2102.941	650	18.733	14	.175	-9.267

-2LL=-2 Log likelihood; LRT=Likelihood Ratio Test; Δdf =Change in degrees of freedom; AIC=Akaike Information Criterion. Comparison model for each measure is always the saturated model. Most parsimonious (best-fitting) model is shown in bold font.

Table 3

Model comparisons for Non-Spatial Alternation (NSA) test

	-2LL	df	LRT	Δdf	p	AIC
<u>NSA Achieving Criterion</u>						
1a Saturated Model	900.398	651	-	-	-	-
2a ACE Model	905.762	654	5.364	3	.147	-0.636
3a ADE Model	905.177	654	4.779	3	.189	-1.221
4a AE Model	905.761	655	5.363	4	.252	-2.637
5a CE Model	907.332	655	6.934	4	.139	-1.066
6a E Model	910.794	656	10.396	5	.065	0.396
<u>NSA Total Errors</u>						
1b Saturated Model	1571.172	642	-	-	-	-
2b ACE Model	1587.389	651	16.216	9	.062	-1.784
3b ADE Model	1587.163	651	15.992	9	.067	-2.009
4b AE Model	1587.388	652	16.216	10	.094	-3.784
5b CE Model	1589.059	652	17.887	10	.057	-2.113
6b E Model	1595.709	653	24.537	11	.011	2.537

-2LL=-2 Log likelihood; LRT=Likelihood Ratio Test; Δdf =Change in degrees of freedom; AIC=Akaike Information Criterion. Comparison model for each measure is always the saturated model. Most parsimonious (best-fitting) model is shown in bold font.

Table 4

Standardized variance components for full and most parsimonious (best-fitting) models

Variance Components				
Model	a ² (95% CI)	c ² (95% CI)	d ² (95% CI)	e ² (95% CI)
<u>Spatial Alternation Achieving Criterion</u>				
ACE	.00 (.00, .00)	.09 (.00, .30)		.91 (.70, 1.00)
E only	-----	-----	-----	1.00
<u>Spatial Alternation Total Errors</u>				
ACE	.00 (.00, .22)	.11 (.00, .23)		.89 (.77, 1.00)
CE	-----	.11 (.00, .23)	-----	.89 (.77, 1.00)
<u>Non-Spatial Alternation Achieving Criterion</u>				
ACE	.25 (.00, .45)	.00 (.00, .29)		.75 (.55, .97)
ADE	.00 (.00, .44)	-----	.28 (.00, .49)	.72 (.51, .95)
AE	.25 (.03, .45)	-----	-----	.75 (.55, .97)
<u>Non-spatial Alternation Total Errors</u>				
ACE	.26 (.00, .43)	.00 (.00, .29)	-----	.74 (.57, .92)
ADE	.16 (.00, .43)	-----	.11 (.00, .46)	.73 (.54, .91)
AE	.26 (.09, .43)	-----	-----	.74 (.57, .91)

CI=Confidence interval; A=Additive genetic variance; C=Common (shared) environmental variance; D=Non-additive/dominance genetic variance; E=Nonshared environmental variance; ACE and ADE=Full models. Models shown in bold font are the most parsimonious models for each variable.