ELSEVIER
JUNE 2014 | VOLUME 55 | ISSN 0010-9452 |
SPECIAL ISSUE:
LANGUAGE, COMPUTERS AND COGNITIVE NEUROSCIENCE
Cortex
A journal devoted to the study of the nervous system and behavior
Guest Editors:
Brita Elvevåg and Peter Garrard
Available online at www.sciencedirect.com
ScienceDirect

Available online at www.sciencedirect.com

ScienceDirect

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

Special issue: Research report

A computational language approach to modeling prose recall in schizophrenia

Mark Rosenstein [a,*], Catherine Diaz-Asper [b], Peter W. Foltz [a,c] and Brita Elvevåg [d,e]

[a] Pearson Knowledge Technologies, Boulder, CO, USA
[b] Clinical Brain Disorders Branch, National Institute of Mental Health/NIH, Bethesda, MD, USA
[c] University of Colorado, Institute of Cognitive Science, Boulder, CO, USA
[d] Psychiatry Research Group, Department of Clinical Medicine, University of Tromsø, Norway
[e] Norwegian Centre for Integrated Care and Telemedicine (NST), University Hospital of North Norway, Tromsø, Norway

ARTICLE INFO

Article history:
Received 4 June 2013
Reviewed 28 August 2013
Revised 11 December 2013
Accepted 22 January 2014
Published online 8 February 2014

Keywords:
Episodic memory
Wechsler memory scale
Text analysis
Latent semantic analysis

ABSTRACT

Many cortical disorders are associated with memory problems. In schizophrenia, verbal memory deficits are a hallmark feature. However, the exact nature of this deficit remains elusive. Modeling aspects of language features used in memory recall have the potential to provide means for measuring these verbal processes. We employ computational language approaches to assess time-varying semantic and sequential properties of prose recall at various retrieval intervals (immediate, 30 min and 24 h later) in patients with schizophrenia, unaffected siblings and healthy unrelated control participants. First, we model the recall data to quantify the degradation of performance with increasing retrieval interval and the effect of diagnosis (i.e., group membership) on performance. Next we model the human scoring of recall performance using an *n*-gram language sequence technique, and then with a semantic feature based on Latent Semantic Analysis. These models show that automated analyses of the recalls can produce scores that accurately mimic human scoring. The final analysis addresses the validity of this approach by ascertaining the ability to predict group membership from models built on the two classes of language features. Taken individually, the semantic feature is most predictive, while a model combining the features improves accuracy of group membership prediction slightly above the semantic feature alone as well as over the human rating approach. We discuss the implications for cognitive neuroscience of such a computational approach in exploring the mechanisms of prose recall.

1. Introduction

Human memory is to a large extent genetically controlled, and thus it is considered to be a heritable, polygenic trait. In schizophrenia impaired cognitive function is a core feature of the illness (Elvevåg & Goldberg, 2000) and some of the most prominent deficits are in verbal episodic memory (Aleman, Hijman, de Haan, & Kahn, 1999; Barch, 2005; Cirillo & Seidman,

* *Corresponding author*. Pearson, 4940 Pearl East Circle, Suite 200, Boulder, CO 80301, USA.
E-mail address: mark.rosenstein@pearson.com (M. Rosenstein).

http://dx.doi.org/10.1016/j.cortex.2014.01.021
0010-9452/© 2014 Elsevier Ltd. All rights reserved.

2003; Kalkstein, Hurford, & Gur, 2010). The disproportionate impairment in verbal episodic memory relative to visual episodic memory may suggest that a useful endophenotype is a deficit in verbal processing, rather than memory impairment *per se* (Skelley, Goldberg, Egan, Weinberger, & Gold, 2008). In this paper, we use recalls from a widely used prose recall test to explore the usefulness of an automated scoring methodology that has the potential to provide equivalent or more sensitive scoring metrics to that of human raters, as well as a more detailed characterization of recall performance over time.[1]

Measures of verbal episodic memory typically include the learning and subsequent recall of word lists or prose passages (stories), and one of the most comprehensive, popular and enduring scales is the Wechsler Memory Scale (WMS; Wechsler, 1945, 1987, 1997, 2009), currently in its 69th year and fourth revision. With minor modifications over time, the Logical Memory subtest has remained a core component of the battery, and is one of the most widely-used measures of prose recall in the research literature (Rabin, Barr, & Burton, 2005).

The Logical Memory task requires participants to repeat back two orally-presented short stories, both immediately after presentation, and following a 30 min delay.[2] The scoring criteria, or rubric, generally specifies that one point is awarded for each key word or narrowly defined concept correctly recalled, with a maximum of 25 points per story, summed for a total raw score out of 50. A measure of forgetting ["percent retained" (Russell, 1988) or "saving score"[3] (Munro Cullum, Butters, Tröster, & Salmon, 1990)] can also be calculated as the total number of items recalled following the delay interval, divided by the total number recalled immediately after initial presentation. Prose recall tasks such as Logical Memory likely rely heavily upon multiple cognitive and memory systems, including language comprehension, conceptual organization, schema formation, working memory, and episodic and semantic memory (Baddeley & Wilson, 2002; Dunn, Almeida, Barclay, Waterreus, & Flicker, 2002). Since performance on this task relies upon hippocampal memory systems (Ho et al., 2006; Lim et al., 2006; O'Driscoll et al., 2001), it is a sensitive assay of verbal episodic memory dysfunction in a variety of neuropsychiatric conditions, including schizophrenia and Alzheimer's disease (Egan et al., 2003; Matsui et al., 2007; Vassos et al., 2010). Importantly, it demonstrates a genetic load effect in schizophrenia, with unaffected siblings typically performing intermediary between patients with schizophrenia and healthy controls (Goldberg et al., 1995; Skelley et al., 2008; Toulopoulou, Rabe-Hesketh, King, Murray, & Morris, 2003). While Logical Memory has proven a useful clinical measure of verbal episodic memory, there are several limitations. Early versions (WMS, Wechsler, 1945; Wechsler Memory Scale-Revised (WMS-R), Wechsler, 1987) relied heavily upon the recall of salient words from the story, known as "story units", yet prose recall is rarely verbatim (e.g., Kintsch, 1998). Rather, it is filled with approximate renderings of the passage that may include substitutions, omissions, additions and elaborations, and shifts in the story's sequence (Lezak, Howieson, & Loring, 2004). These common deviations in recall are not adequately captured by the relatively simplistic "story units" measurement. More recent revisions of the test (e.g., WMS-III, Wechsler, 1997) have introduced "thematic" scoring units in addition to story units, wherein larger chunks of discourse pertaining to a theme are sought rather than the verbatim recall of select key words, presumably to better capture gist recollection. However, Dunn et al. (2002) contend this measure is merely a subset of story units and adds no additional information. The approach further relies on the subjective judgment by the scorer about the degree of match of recall to themes. For these reasons, in this study only the story unit rubric was used.

A few studies illustrate how departing from the constraints of standard administration and scoring can provide complimentary information on verbal episodic memory function. For example, when Skelley et al. (2008) examined episodic memory function in patients with schizophrenia, their unaffected siblings and healthy unrelated controls, they utilized the "savings score" calculation on total raw scores on Logical Memory at three different time points (immediate, 30 min, and 24 h). They reported that both patients and siblings displayed the greatest impairment in initial learning (from immediate to 30 min) and little impairment in long-delay savings (from 30 min to 24 h).

An alternative approach to obtain further information from prose recalls is to assay the effect of the underlying cognitive processes integral to prose recall – the sequential construction of the words and semantic processes – but this approach may introduce a level of subjectivity potentially compromising reliability and validity (Dunn et al., 2002). However, a way to obviate this concern is to employ automated language analysis methods. The first question we address is whether automated methods can perform as well as humans in the scoring task, and then having established a baseline performance, whether features arising from automated analysis might actually outperform the existing rubric in predicting group membership (i.e., diagnosis). We have previously shown that departing from global scoring techniques and employing a data-driven methodology can provide useful information concerning cognitive strategies that individuals use in order to remember lists of words (Longenecker, Kohn, et al., 2010). In the case of prose recall, given the "story unit" rubric's strong emphasis on capturing exact words and phrases,[4] a language sequential categorization algorithm based on natural language processing techniques may be able to capture much of how humans

[1] Although we illustrate this method with a test from the WMS-R (Wechsler, 1987), the techniques can naturally be applied to other verbal memory tests.

[2] A third recall at 24 h was added to the protocol for this study.

[3] For clinical purposes, the raw score may be converted to a standardized scaled score (0–19) based on the normative tables published in the test manual. The concept of "saving score" has a long history (e.g., Ebbinghaus, 1885/1913). Robinson and Heron (1922) define "saving score" in the context of memorizing lists, though in their case practice over time improved performance, so instead of directly reporting the fraction presented here, the fraction was first subtracted from 100. This metric is reportedly less vulnerable than standardized scaled scores to the well-documented declines in performance on the Logical Memory test with advancing age (Lezak et al., 2004), and also differentiates cortical from subcortical dementias (Tröster et al., 1993).

[4] Especially in the WMS-R which is the test version we employed.

score this task. In a number of text categorization tasks, the language sequence scoring method based on a text's character *n-gram* frequency profile (Cavnar & Trenkle, 1994) has been successfully applied. The next two techniques we employed attempt to measure more general characterizations of the recalls than specified by the rubric. The first technique again uses the character *n-gram* frequency, but in this case compares the recall to the expected frequency of standard English. The second technique, Latent Semantic Analysis (LSA), has recently been employed to better characterize performance on the Logical Memory task (Dunn et al., 2002; Lautenschlager, Dunn, Bonney, Flicker, & Almeida, 2006) by allowing semantic comparisons at a meaning (thematic) level. In brief, LSA is an automated mathematical procedure that uses corpus-based information to perform semantic comparisons on words and units of text (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). Due in part to its automaticity and consistency, LSA has the advantage over traditional scoring methods of not only being objective and reliable (Landauer et al., 1998) but also potentially more sensitive to elements of recall missed by standard scoring methods (Dunn et al., 2002). Indeed, in a proof of concept study, Dunn et al. (2002) demonstrated the utility of LSA as an alternative to standard scoring on the WMS-III version of Logical Memory (Wechsler, 1997) by contrasting the two scoring systems in groups of both cognitively-intact and impaired older individuals. They used LSA to measure the similarity of recall to the original text by calculating the cosine of the angle between a recall attempt and the original text. They reported that LSA was at least as valid and sensitive in detecting an effect of cognitive impairment, demonstrating that the three variables (LSA, thematic scoring units and story units) were highly correlated for both immediate and delayed recall of the Logical Memory stories, with correlations ranging from .69 to .94.

We sought to examine the sensitivity of measurement of these computational sequential and semantic similarity metrics on prose recall and how they might modulate performance at increasing retrieval intervals and as a function of diagnosis (namely in patients with schizophrenia, unaffected siblings and unrelated healthy control participants).

2. Methods

2.1. Participants

Patients with schizophrenia ($N = 28$), unaffected siblings ($N = 18$), and unrelated healthy control participants ($N = 76$) between the ages of 18 and 60 years with an estimated premorbid IQ greater than 70 were included. All participants were recruited as part of the Clinical Brain Disorders (NIMH) Schizophrenia Sibling Study (D.R. Weinberger, PI) (Egan et al., 2000), and completed a battery of neuropsychological tests assessing multiple cognitive domains. The results of these tests are not reported here, aside from two tests used to index current intellectual function (an abbreviated form of the Wechsler Adult Intelligence Scale-Revised (WAIS-R); Missar, Gold, & Goldberg, 1994) and an estimate of premorbid intellectual function (Wide Range Achievement Test-Revised (WRAT-R); Jastak & Wilkinson, 1984; Wiens, Bryan, & Crossen, 1993). All participants provided written informed consent to participate (according to the NIMH Institutional Review Board guidelines and the regulations and ethical guidelines of the National Institutes of Health Office of Human Subjects Research) in the NIMH Schizophrenia Sibling Study, which is an ongoing U.S. investigation of neurobiological abnormalities related to the genetic risk for schizophrenia. All participants were screened by two board-certified psychiatrists using semi-structured psychiatric interviews, third-party informants, toxicology screening, and cognitive testing exclusions previously described (Egan et al., 2000, 2001). Participants were also excluded if they had a history of significant medical or neurological disorders, such as epilepsy or traumatic brain injury. All patients met DSM-IV criteria (First, Spitzer, Gibbon, & Williams, 1997) for schizophrenia or schizoaffective disorder, depressive type, and 68% were on antipsychotic medication at the time of study. Psychoactive medications influence verbal memory (e.g., Baitz et al., 2012; Brébion, Bressan, Amador, Malaspina, & Gorman, 2004; Mori, Nagao, Yamashita, Morinobu, & Yamawaki, 2004), but further considering or controlling for these effects in this context was neither a goal nor practical due to sample size for this study. All siblings were free from schizophrenia spectrum disorders. All control participants were free from DSM-IV lifetime psychiatric illness or current substance abuse.

Groups did not differ significantly in age at time of testing (see Table 1 for demographic data). However, there were group differences in gender distribution [$X^2(2, N = 122) = 6.47$, $p = .04$] (and indeed women are often underrepresented in research in schizophrenia – see Longenecker, Genderson, et al., 2010). As is typically reported when comparing patients with schizophrenia to healthy controls (Weickert et al., 2000), groups differed significantly in terms of educational attainment [$F(2,121) = 10.04$, $p < .001$] and current IQ [$F(2,121) = 17.69$, $p < .001$]. Post-hoc contrasts revealed that for both education and current IQ, patients exhibited significantly lower levels than both siblings and controls ($p < .05$), but that siblings and controls did not differ from each other. Between-group differences were also seen in terms of a measure of estimated premorbid IQ [$F(2,121) = 5.07$, $p < .01$] with patients exhibiting significantly lower levels than siblings ($p < .05$) and controls, the latter two whom did not differ from one another.

2.2. Prose recall, transcript preparation and inter-rater reliability

All participants completed a test of episodic memory function (the Logical Memory subtest of the WMS-R – Wechsler, 1987). As noted above, the Logical Memory test consists of two brief stories read to the participant by an examiner. The participant is asked to recall as much of the story as they can (*immediate recall*), and following a delay of 30 min the participant is again asked to recall as much of each story as possible (as a measure of *short-delay recall*). We also added a 24 h delayed recall condition (a measure of *long-delay recall*). Following immediate encoding, participants were told that they would be asked about the task again later (30 min recall); in contrast, no

Table 1 – Demographic data for patients, siblings and healthy controls.

	Patients ($n = 28$)	Siblings ($n = 18$)	Controls ($n = 76$)	p-Value[a]
Age (years)	30.82 (±9.19)	33.61 (±11.81)	32.25 (±9.80)	.642
Education (years)	13.86 (±2.32)	15.94 (±1.59)	16.11 (±2.43)	<.001
Gender (M/F)	20/8	7/11	35/41	.04
Current IQ (WAIS-R)	89.06 (±13.62)	112.44 (±8.76)	106.38 (±16.44)	<.001
Estimated "premorbid" IQ (WRAT-R)	101.30 (±12.04)	110.89 (±6.36)	107.93 (±11.55)	.008
CPZE[b]	559.85 (±660.48)[c]	–	–	–

[a] ANOVAs for all continuous variables, and chi-square for gender.
[b] CPZE: Chlorpromazine equivalents.
[c] Range 0–2700.

warning was provided about the long-delay (24 h) recall task. Dependent variables were raw recall scores at each time point, for each story (max = 25 points each) and combined (max = 50 points), as well as savings scores for the short (30 min) delay (immediate to 30 min) and long delay (30 min to 24 h) intervals.

We generated two streams of scores. First, as each recall was spoken by a participant, the audio of the recall was recorded and WMS-R scores were generated by the experimenters for each story in real-time. These individual story unit scores were summed for each participant at each recall-time, and the individual story unit scores were not retained. We refer to this summed score (max = 50) as the original combined score. The recorded audio from the recalls was also transcribed. For the automated text analysis of the transcripts, only the content of the participant recall was used. All transcriber meta-comments [such as "(pause)" or "(equipment noise)"], any experimenter speech including experimenter meta-data (such as "Participant number one hundred. Um. Logical Memory Immediate recall") and any experimenter prompting of the participant (such as "Anything else you remember from that story?") were excluded from the transcripts. The final data set included the cleaned transcripts for 353 recalls for each of Story 1 and Story 2 with an additional 353 original summed scores.

We chose to rescore the transcripts to allow analysis at the individual story level and to allow the production of human scores that were blind to group membership (i.e., patients, unaffected siblings, unrelated controls) and recall-time (i.e., immediate, 30 min or 24 h later). With the original summed scores, the experimenters had access to a myriad of information sources implicit in a face-to-face setting, which goes well beyond the text of the recall. Due to issues of subjectivity and evidence that thematic units provide no additional information (Dunn et al., 2002), only the story unit rubric was considered. Two human raters were recruited[5] to provide WMS-R scores from the transcribed recalls to provide a comparable condition to that faced by automated analysis. Scorer 1 was presented with all the recalls from Story 1 in a random order followed by all the recalls from Story 2 also in a random order. Scoring proceeded by entering scores for each participant at each recall-time into a spreadsheet version of the standard WMS-R story unit form. For Scorer 2, who served as one of the sources to measure inter-rater reliability, a stratified random sample was taken of six recalls at each combination of recall-time × group for a total of 54 recalls for each story. The correlation between the score for Scorer 1 and Scorer 2 for Story 1 was $r(52) = .99$, $p < .001$ and the correlation for Story 2 was $r(52) = .97$, $p < .001$. Since scores are typically reported as the sum of scores from the stories, the correlation between the summed scores from Scorer 1 and Scorer 2 was $r(52) = .99$, $p < .001$. Comparing the original summed scores to the summed scores for Scorer 1 yielded a correlation of $r(351) = .98$, $p < .001$ and for Scorer 2 $r(52) = .98$, $p < .001$ (with $n = 353$ and $n = 54$ respectively). Since the correlations between the blind and original summed scores are practically identical, concerns over bias in this case under the story unit rubric do not seem justified.

2.3. Computational language features

2.3.1. Character n-gram

The story unit rubric, by focusing almost exclusively at the level of exact key word recall in distinction to paraphrases or recalling main ideas, may allow automated scoring via methods based on sequential language order. Character and word sequences capture aspects of word choice, syntactical word ordering as well as language fluency and grammatical flow. One of the most parsimonious, though quite powerful syntax scoring methods is based on a text's character *n-gram* frequency profile (Cavnar & Trenkle, 1994). This measure compares character patterns between texts.[6] Briefly, *n-grams* are segments of length n drawn from a text. The unit of analysis of a text string can be at the level of characters or words, with unigrams (or equivalently 1-gram) being all the individual components of that string, whereas, 5-grams encompass all combinations of five characters in a row encompassing the flow of one word, its punctuation or spaces and the next word. The counts of n-grams of a text typically follow a Zipf distribution (Zipf, 1935). Cavnar and Trenkle's insight was that a reasonably small portion of the

[5] One rater had a PhD and 12 years of relevant experience, and the other rater had a Masters and 5 years of relevant experience.

[6] The technique of Cavnar and Trenkle (1994) should not be confused with *n-gram* statistical language models (see e.g., Jurafsky & Martin, 2009; Chapter 4), in that their categorization technique requires relatively small data structures and is quite fast.

distribution representing the most frequent *n*-grams could be used as a text's profile (they used the 300 most frequent *n*-grams) and then appropriately defining a metric over pairs of profiles allows the production of a sequential similarity measure.

In our current study, a profile for each of the two original stories was generated and then the distances from the profiles of each recall to the profile of the original story computed. Our expectation is that the less well a recall profile matches the profile of the original story (a larger distance) the lower the human recall score should be. How well the profile distances (negatively) correlate to the human scores will provide a measure of the accuracy of this automated measure. A second method in which the *n*-gram profile measure can assess text is via a model of the "English likeness" of the *n*-gram sequences of a text. For this measure, an English profile built on a publically available English corpus is constructed and then compared to the profiles of the recalls. Notice that this measure is independent of the original story and is entirely based on how similar the recall is to standard English as encoded in the English profile. Our expectation is that the less well a recall profile matches the profile of "standard English" (the greater its profile distance to the English profile), the lower should be the human scoring for the recall. It is possible that with an increasing severity level in the clinical presentation of the illness, the further the recall will drift from standard English, and that siblings will exhibit an intermediate level of closeness to standard English. It is also possible that this feature would be useful in measuring progression of a disease (such as dementia) where language increasingly deviates from the 'norm'. The information theory and statistics literature contain numerous measures to compare frequency distributions (see Jurafsky & Martin, 2009). In previous unpublished work, we found that using the Kullback–Leibler divergence (Kullback & Leibler, 1951) as the comparison metric outperforms the "out-of-place" distance metric of Cavnar and Trenkle and the Kullback-Leibler metric is applied using *textcat* (Hornik, Rauch, Buchta, & Feinerer, 2012). Although this metric is not strictly speaking a "distance" (for instance Kullback–Leibler is not symmetric), in this context the metrics behave intuitively like distances, so we refer to the output of the metric as a distance. Since we wished to potentially include word level frequency comparisons, character *n*-grams of sizes ranging from 1 to 5 were used. The "English like" profile was built from the English corpus provided as part of the European Corpus Initiative Multilingual Corpus I (ECI/MCI) (Armstrong-Warwick, Thompson, McKelvie, & Petitpierre, 1994).

2.3.2. *LSA*

Across many task domains, LSA (Deerwester et al., 1990) has been shown to capture semantics in ways that can be usefully applied in similarity comparisons. LSA is a corpus-based statistical modeling method based on computing a reduced dimension singular value decomposition of a reference corpus (see Landauer et al., 1998 for a technical description). Vectors in this reduced dimensional "semantic space" represent words and text passages. Computing the cosine between two vectors in this space generates a semantic similarity measure, which can be computed even if the two units of text share no words in common. This technique has been widely used in such fields as information retrieval (Deerwester et al., 1990), automated essay scoring (Foltz, Laham, & Landauer, 1999), analyzing prose recall (Dunn et al., 2002), and in the analysis of prose from patients with schizophrenia (Cabana, Valle-Lisboa, Elvevåg, & Mizraji, 2011; Elvevåg, Foltz, Weinberger, & Goldberg, 2007; Elvevåg, Foltz, Rosenstein, & DeLisi, 2010).

For our current study, vectors representing the original story were compared with vectors from each of the recalls. The expectation was that less semantically similar recalls would receive lower human recall scores. Our LSA analysis was conducted using our own LSA software. The semantic space was built from the TASA corpus (Zeno, Ivens, Millard, & Duvvuri, 1995) with dimension reduction to 300 dimensions.

2.4. *Analysis approach*

To analyze the performance of the sequential and semantic features, we used two statistical modeling techniques, linear mixed-effects models and proportional odds logistic regression. Although these models are less widely used in the analysis of recall data, we argue below that they provide a more unified view of the data and better capture its underlying structure than some of the existing analysis methods, especially with data based on repeated measures and with potential individual and family effects.

2.4.1. *Statistical method – linear mixed-effects model*

To accurately tease out the effect of group membership and recall-time on the recall score, it is critical to recognize and attend to the covariance structure of the data. There is potentially both an individual effect exposed through a repeated measures design and a family component. That these data cannot be treated as independent observations is clearly demonstrated in Table 2, which shows the correlations between scores measured at different recall-times for Story 1 and Story 2 for the control group. We chose the control group to avoid confounding individual with group correlation, but this pattern of correlations is nearly identical for the entire data set (though slightly higher than evidenced by the control group). The smallest correlation for the control group is .85 and all correlations are significant with $p < .001$. These results clearly indicate a strong individual effect in this task, which unless handled correctly

Table 2 – Correlation of human recall scores at different recall-times for control group.

	Immed. to 30 min	Immed. to 24 h	30 min to 24 h
Story 1	.90	.86	.94
Story 2	.85	.90	.93

can distort standard error estimates and possibly bias parameter estimates.

A linear mixed-effects model was used to describe the relation between human recall score as response with group and recall-time as explanatory fixed effects. In addition, the model provides a random effect for individuals and a random effect for families. The model further allows the conservation of observations in that observations at all three recall-times are not required, so participants with some missing observations can still be included in the analysis. An advantage of this approach over for example the one taken in Skelley et al. (2008) is the ability to directly estimate effects. A disadvantage is that it will not directly estimate "saving scores". Instead, additive differences (the effects) for group and recall-time are estimated including potential interactions.

To fix notation, Fig. 1 shows the potentially most complex model considered. Y is the response, the human score, which is assumed to be generated by the fixed effects (subscripted "b"), the random effects (subscripted "z") and an error term e. The fixed effects are estimated, as are the standard deviation for the random effects, which are constrained to have means of zero. The control group is the baseline level for the group factor, and immediate recall serves as the baseline level for the recall-time factor.

All modeling was conducted using the lme4 package (Bates, Maechler, & Bolker, 2012) in the R statistics environment (R Core Team, 2012). The results of modeling comprise estimates for the standard deviation of the random effects as well as parameter estimates for the fixed effects and their standard errors. The confidence intervals for these estimates are typically not symmetric, so in addition to reporting *t*-values, we profile the likelihood to obtain a 95% confidence interval for the standard deviations and parameter estimates. If the 95% confidence interval does not include zero, we conclude the estimate is significant at the .05 level. From an estimated model, the random effects provide information about how much variability is attributable to individual differences, while the fixed effects will account for the impact of the levels of group and recall-time, as well as any effects of their interactions.

$$Y_{it} = b_0 + b_g + b_t + b_{gt} + z_i + z_f + e_i$$

where individual i is a member of group g and belongs to family f

Y_{it} – the response, the recall score for individual i at recall-time t

b_0 – the intercept, in this case the expected score at immediate time and control group

b_g – the effect on recall of group g with respect to control group

b_t – the effect on recall score at recall-time t with respect to immediate recall-time

b_{gt} – interaction term between group g at time t

z_i – a random effect for individual i

z_f – a random effect for the family f of individual i

e_i – the residual error

Fig. 1 – Specification of a linear mixed-effects model of recall score predicted by group and time.

The model selection strategy starts with a base model and adds features, so that a series of nested models are generated. This allows the use of a likelihood ratio test to determine if the improved model is worth the added complexity. We also present the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978), which are both model selection criteria that utilize model likelihood, but penalize model complexity, with BIC penalizing complexity more heavily. With both criteria, a lower value is better.

2.4.2. *Statistical method – proportional odds logistic regression*

In the case of diagnosing schizophrenia, there is currently no single biological or neurocognitive test that will alone or conclusively confirm a diagnosis of schizophrenia. Rather it is a descriptive process based on many aspects derived from the clinical examination. Thus, in terms of the current study the 'gold standard' for any automated classification is how the patients were originally classified clinically (i.e., without the help of extensive analysis of speech). Therefore, for current purposes we have chosen to use the ability to predict group membership in evaluating the usefulness of the language features. This appraisal requires the use of a statistical classification technique. For classification tasks involving only two groups, logistic regression (e.g., see Agresti, 2007) is a common and powerful technique to estimate the group membership probability. For this data with its three ordered groups, a cumulative probability model, where the probability of being in a given group increases in an ordinal fashion from patient to sibling to control as the "closeness" of the recall to the stimulus as measured by the features increases.[7] Concretely, we specify a proportional odds logistic regression model (McCullagh, 1980) with group category as the response and recall-time and various features as the explanatory variables.

The procedure we followed used a standard proportional odds estimation, to compare how well the different models predict group membership. The measure that best predicts group membership is the one that better captures the putative "schizophrenia signal" in the text. However, since the covariance structure is not being accounted for, we need to be cautious in interpreting standard errors and estimates from the models.

For this analysis, we used the polr function in the MASS package (Venables & Ripley, 2002) in the R statistics environment (R Core Team, 2012). Since the confidence interval for proportional odds models may not be symmetric, we present a confidence interval based on the profile of the log-likelihood function (Venables & Ripley, 2002). While the modeling is in terms of odds (or more accurately log of odds), our discussion will be primarily in terms of group membership probability. As with previous modeling, the model selection strategy starts with a simple model, adds features and then uses the likelihood ratio test along with AIC and BIC to base decisions on whether improved prediction justifies the added complexity.

[7] For character *n*-gram profile, which is a distance instead of a similarity, the order of groups is reversed, hence our use of "closeness".

3. Results

First we present the analysis with traditional measures of prose recall, and then present the analysis of the computational approach.

3.1. *Traditional human scored measures of prose recall*

3.1.1. *Immediate and delayed recall performance as a function of group*

As shown in Fig. 2, between-group differences were seen across all three time points in Logical Memory recall total raw score [immediate: $F(2,121) = 24.31$, $p < .001$; 30 min delay: $F(2,121) = 26.80$, $p < .001$; 24 h delay: $F(2,121) = 28.05$, $p < .001$]. Performance of the patient group was significantly lower on raw recall scores than both healthy controls and the sibling group across all three groups [patients *vs* controls: $t(102) = -6.60, -7.02, -7.17$ for immediate, 30 min and 24 h recall respectively; patients *vs* siblings[8]: $t(44) = -4.23, -4.37, -4.63$ for immediate, 30 min and 24 h recall respectively; all $p < .0001$]. In contrast to patients, the sibling group did not differ from controls across any of the three recall points [$t(92) = -1.24, -1.29, -1.12$ for immediate, 30 min and 24 h recall respectively].

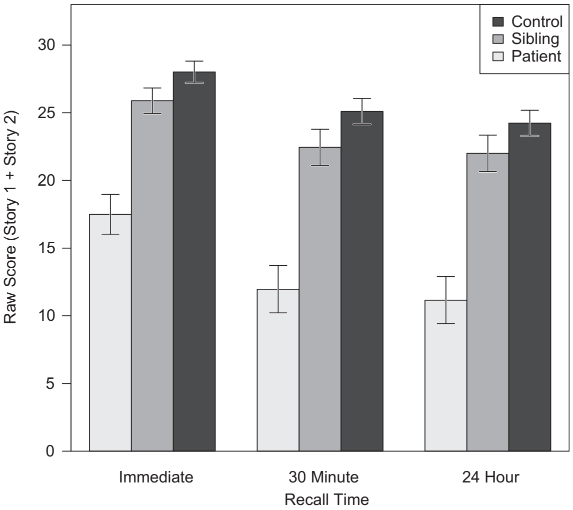

Fig. 2 – Mean recall performance on combined Story 1 and Story 2 (maximum = 50 points) of the Logical Memory subtest of the WMS-R, as a function of time and group. Error bars represent standard error.

3.1.2. *Short and long-delay savings as a function of group*

As seen in Fig. 3, between-group differences were found in short-delay savings scores (30 min recall/immediate recall × 100) on the Logical Memory task [$F(2,121) = 20.67$, $p < .0001$], but not in long-delay savings scores [24 h recall/30 min recall × 100; $F(2,121) = .33$, $p = .72$]. In the patient group, short-delay savings scores were significantly lower relative to both controls [$t(102) = 6.08$, $p < .0001$] and siblings [$t(44) = 3.84$, $p < .001$], who did not differ from each other.

3.2. *Modeling human score*

3.2.1. *Modeling human score as a function of group and recall-time*

Equation (1) is the initial model, which is based on our understanding of the literature indicating that both group and recall-time are important determinates of overall recall score. The model also includes a random effect for individual ability on this task. We first elaborate this model and then compare the more complex models with each simpler model to see if the added complexity significantly improves the model.

$$Y_{it} = b_0 + b_g + b_t + z_i + e_i \tag{1}$$

The first alternative model was designed to gauge whether a family random effect improved the model. Since there are significant issues with the family composition of the data, these results are delegated to the Appendix. Briefly, the results do not indicate that adding a family effect significantly improved the model for human scores for either Story 1 or Story 2, though there was just a hint of family cohesion in modeling LSA cosine scores, results which are all elaborated in the Appendix. The fixed effects estimates for equation (1) are given in Table A2 and the random effects for Story 1 are presented in Table A3. The lower bound for the 95% confidence interval for the standard deviation of the family effect is

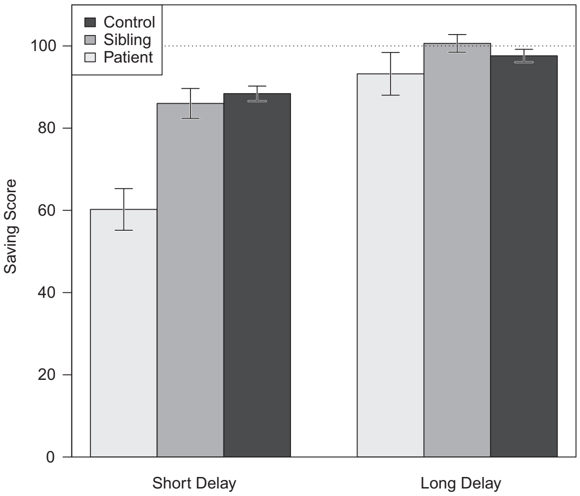

Fig. 3 – Saving scores for the Logical Memory subtest of the WMS-R, as a function of group. Short delay is computed as mean 30 min score/immediate score and long delay is mean 24 h score/30 min score. Error bars represent standard error.

[8] This analysis follows Skelley et al. (2008), but as noted in the description of this data, the sibling and patient groups are not independent samples, making the *t*-test standard errors potentially problematic.

undefined (the equivalent of the confidence interval containing zero, for a parameter that is not defined for values less than zero), so we dropped the family effect as not statistically significant.

The next alternative model considered adding interaction terms. Table 3 shows the estimates for the random effects for Stories 1 and 2 for the model including interaction terms. The standard deviation for the individual effect (labeled ID in the Groups column) is about 4 in both models and the 95% confidence intervals for the estimates are well away from zero, so both are statistically significant. Since the potential score range is 0–25, with an observed range of 0–21 for Story 1 and 0–22 for Story 2 in this data, the range of plus or minus one standard deviation of the individual random effect covers a bit more than a third of the score range. This strengthens the recall-time correlation evidence indicating that individual effects are a significant proportion of performance. The magnitude of this individual variability should be kept in mind as we examine the fixed effects in Table 4. The standard deviation of the residual distribution (the variation not explained by the model) presented in the rows labeled Residual in Table 3 is 1.48 in Story 1 and 1.40 in Story 2.

Table 4 shows the estimates, standard errors, and 95% confidence intervals for the fixed effects for Story 1 and Story 2. The baseline for the contrasts is the control group and immediate time-recall. For Story 1 all the main effects are significant as are the two patient × recall-time interactions. Interactions are shown in rows labeled with group × recall-time. The sibling group and recall-time interactions are not significant, though the likelihood ratio test result shown in Supplemental information (Table S1) strongly rejects that the models have equal predictive power ($p = .0074$). The AIC concurs with a decrease from 1676.70 to 1670.70 for the more complex model, while BIC increased. We chose to examine the more complex interaction model as the best description of the data generating process, which we will now examine in detail.

One interpretation of the Story 1 model is to start with an average participant and then note the changes as group and recall-time are varied. This model indicates that on average an individual will have a recall score of 14.84. That person draws from a normal distribution with mean zero and standard deviation of 4.06 and the drawn value, their individual ability, is added to the starting mean score of 14.84. If that person is in the control group (the baseline), their recall score is this value, if they are in the sibling group, their score drops on average 2.22, and if in the patient group their score drops on average 5.73. For immediate recall (the baseline), there is no change in the recall score, while at 30 min there is a 2.42 drop and at 24 h there is a 2.75 drop. There are statistically significant interactions for the patient group with recall-time, so there is an additional penalty for patients added to the previous drops at the 30 min recall of 1.07 and at 24 h of 1.20. This analysis is comparable to the saving score analysis reported by Skelley et al. (2008), where the statistically significant drop of 5.73 plus the additional interaction drops strongly indicates patients recall less than controls. Fig. 4 plots both the mean values from the data, and the model predictions. The interaction effect as evidenced by differing slopes per group for patient versus control (with patients' slope being steeper than controls) is quite evident especially between immediate and 30 min recall. While the differences between the model predictions and the data means are all small, in fact less than a score point, the model allows quantifying the interaction effects, isolating and estimating the standard deviation of the individual ability distribution, and, at least for this data, to be able to rule out a family effect all with more accurate confidence interval estimates.

Table 4 – Fixed effects for interaction models for Stories 1 and 2.

	Estimate	Std. err.	t Value	2.5%	97.5%
Story 1					
Fixed effects:					
Intercept	14.84	.50	29.87	13.87	15.80
Sibling	−2.22	1.13	−1.96	−4.43	−.02
Patient	−5.73	.96	−5.99	−7.59	−3.86
30 min	−2.42	.24	−9.91	−2.89	−1.94
24 h	−2.75	.25	−11.15	−3.23	−2.27
Sibling × 30 min.	.86	.55	1.57	−.20	1.93
Patient × 30 min.	−1.07	.47	−2.28	−1.99	−.16
Sibling × 24 h	.88	.56	1.57	−.21	1.97
Patient × 24 h	−1.20	.48	−2.49	−2.13	−.26
Story 2					
Fixed effects:					
Intercept	13.31	.48	27.98	12.38	14.24
Sibling	−.48	1.08	−.44	−2.59	1.64
Patient	−4.95	.92	−5.41	−6.74	−3.17
30 min	−1.00	.23	−4.33	−1.45	−.55
24 h	−.94	.23	−4.05	−1.40	−.49
Sibling × 30 min.	.33	.52	.64	−.68	1.35
Patient × 30 min.	−.94	.44	−2.12	−1.81	−.08
Sibling × 24 h	−.22	.53	−.41	−1.25	.81
Patient × 24 h	−1.43	.46	−3.13	−2.31	−.54

Table 3 – Random effects for interaction models for Stories 1 and 2.

Groups	Name	Var.	Std. dev.	2.5%	97.5%
Story 1					
Random effects:					
ID	Intercept	16.50	4.06	3.53	4.60
Residual		2.18	1.48	1.34	1.60
Number of obs: 353; ID: 122					
Story 2					
Random effects:					
ID	Intercept	15.19	3.90	3.39	4.41
Residual		1.96	1.40	1.26	1.52
Number of obs: 353; ID: 122					

For Story 2, the main effect on the sibling group is not significant, so in this case there is no statistically significant penalty between the control and sibling groups. The mean score for the control group and immediate recall is 13.31, which is over a point below the value for Story 1. The patient drop is 4.95 and the 30 min recall drop is 1.00 and 24 h is .94, all below the Story 1 values. Except for the absence of a significant sibling main effect, the relationships are very similar to

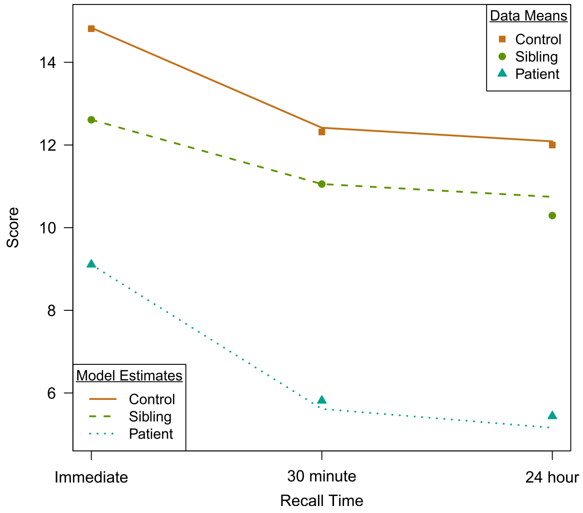

Fig. 4 – Mean human scores and model predictions for Story 1.

Story 1. To statistically compare the relation between the models for Story 1 and Story 2 requires explicitly representing the story in the model to see if they were significantly different. With this model of how group and recall-time affect the human score, we next turn to automatically predict the human score from computational language features of the recalls.

3.2.2. *Modeling human score using language sequence features*

We first examine how well a character *n*-gram sequential feature can predict human scores.

Table 5 contains descriptive statistics for the word and character counts for the two stories and averages for the recalls. Both stories contained the same number of words, while Story 2 contains about 10% more characters, which is reflected in Story 1 having 4.09 characters per word, while Story 2 has 4.46 characters per word. Selecting *n*-gram sizes from 1 to 5 captures most of the information from the average sized word.

For each of the Story 1 recalls the Kullback–Leibler distance to Story 1 was generated and also to the reference English profile. For each Story 2 recall the distance to Story 2 and the English profile were generated[9] (Additional details in Supplementary information *n*-gram analysis).

Table 6 shows the correlations between the human scores and the sequence distances. Since the character *n*-gram sequential measure is a distance, it gets larger the further profiles are apart, which explains the negative correlation with the human ratings. The first striking result from Table 6 is that the distance between the recall profile and the story profile and human ratings of prose recall are very highly correlated at −.92 for Story 1 and −.93 for Story 2. This measure captures a large fraction of the human score variation, indicating that much of the human scores could be automatically scored. These high correlations suggest that the sequential measure performs very similarly to the human scorers. The *n*-gram "English-like" measures have lower correlation to the human scores, with correlations for Story 1 of −.69 and for Story 2 −.65. While these correlations are lower than the other type of recall measure, this does not preclude this measure accounting for other differences in the recalls not captured by the story unit rubric. Since the correlations of the distance between the recall profile and the story profile are quite close to human performance, we also explored using regression to put the profile distance on the same scale as the human scores to explore agreement (see Supplementary information Additional Measures).

Table 5 – Summary of text statistics for stories and recalls.

Type	Story	Mean chars (SD)	Mean words (SD)	Mean chars/word (SD)
Story	1	278	68	4.09
Story	2	303	68	4.46
Recalls	1	203.81 (81.45)	52.59 (21.29)	3.88 (.28)
Recalls	2	225.49 (82.86)	56.58 (21.28)	4.00 (.31)

3.2.3. *Modeling human score using semantic features*

LSA vectors were computed for each of Story 1 and Story 2 and for all the recalls. To represent the semantic distance of a recall to the original story, the cosine between the vector for the source story and the recall was computed. The correlation between the scores Scorer 1 adjudged for Story 1 recalls to the cosine between the text vector for Story 1 and the Story 1 recalls was $r(351) = .83$, $p < .001$ and for Story 2 was $r(351) = .79$, $p < .001$. For Scorer 2 the correlation for Story 1 recalls was $r(52) = .84$, $p < .001$ and for Story 2 was $r(52) = .80$, $p < .001$. Since our goal in computing semantic similarity was to capture information beyond that specified in the rubric, it was not unexpected that just as the sequence feature scores did not fit human scores exactly for the "English-like" measure, it is also not surprising given the emphasis of the story unit rubric on key word matching that these correlations are below those for sequence. What is more interesting in is how well these different features predict group membership (i.e., diagnosis). We examine this question next.

3.3. *Comparing features by predicting group membership*

The previous analysis examined how well automated measures could replicate human scores.

However, a significant assumption underlying that approach is that the existing rubric underpinning the scoring of the recalls optimally uses the information available within the recalls. In this last set of analyses, we examine that claim. As mentioned earlier, in the case of diagnosing schizophrenia the gold standard is the clinical

[9] We also generated profiles using the default metric from Cavnar and Trenkle's (1994) paper, but those distances did not perform as well so are not presented.

Table 6 – Correlation between human and sequential measures of recalls.

	Profile recall 1 to Story 1	Profile recall 1 to English	Profile recall 2 to Story 2	Profile recall 2 to English	Human score Story 1	Human score Story 2
Profile recall 1 to Story 1	1.00	.78	.72	.49	−.92	−.69
Profile recall 1 to English	.78	1.00	.62	.61	−.69	−.56
Profile recall 2 to Story 2	.72	.62	1.00	.74	−.70	−.93
Profile recall 2 to English	.49	.61	.74	1.00	−.46	−.65
Human score Story 1	−.92	−.69	−.70	−.46	1.00	.71
Human score Story 2	−.69	−.56	−.93	−.65	.71	1.00

exam. For our current purpose a continuous metric (e.g., severity of illness) would provide stronger evidence in a comparison among the human scores, the LSA cosine similarity measures and the character *n*-gram profiles to that standard, but diagnosis is currently the best measure available. Thus the analysis evaluates whether these new measures can be as, or more sensitive than the human ratings of recall. What this data set does provide is the categorical variable that accurately distinguishes among the three groups, namely of controls, siblings, and patients. We model this categorical variable on a single dimension representing the probability of group membership and the performance on predicting group will allow a comparison across the features. Before explicit modeling, we first visualize the proportions of group membership as the features vary.

The resulting change in the group membership probabilities as a feature varies allows investigating the ability of a feature to utilize characteristics of the recall text as a link to schizophrenia. Fig. 5 shows conditional density plots (Hofmann & Theus, 2005) for each of the features under consideration. This type of plot reveals from the data the varying proportion in each group as the feature value is varied. Each row presents a single feature chosen from the set: the human score, the LSA cosine distance and the character *n*-gram profile, where the left column presents recalls from Story 1 and the right column from Story 2.

The conditional density plot applies a kernel density function to smooth the group proportions to generate these plots. Shading indicates group membership with control being darkest to patient being lightest. Panel A shows the conditional density plot for human score on Story 1. For participants achieving the lowest score, the group with the largest proportion is patients, the next largest is controls and the remainder consists of siblings. As score increases the proportion of controls and siblings increases. At the highest levels of score, the patient proportion drops to nearly zero and the sibling proportion has decreased from its peak size achieved at intermediate score levels. These plots are a visual expression of the expected outcome that at low scores patients will dominate, while at high scores controls will dominate and most siblings will be located along a range of intermediate scores.

This overall pattern is repeated in all the panels, but there are noticeable differences as well. Panel B shows human scores for Story 2. Unlike the other panels, the increase in control membership with increase in score atypically does not exhibit a mostly monotonic increase, but instead displays a plateau for controls at intermediate scores. The “island” of siblings at intermediate scores is present as it is in most of the other plots. In Panel B, there is a decrease in proportion of control membership at the very highest scores, which reflects two patients scoring quite well (immediate recall above 20) on Story 2, who likely regressed toward the mean in Story 1. (With a larger sample, it might be revealed that Story 2 is not quite as discriminating as Story 1.) Panels C and D show the conditional density plots for the LSA cosine similarity for Stories 1 and 2. For very low values of cosine, the membership in the sibling group is empty, allowing a clean separation between groups. A similar effect, but evidenced at larger distances is seen in Panel E for the character *n*-gram profile measure for Story 1. The plots in Panels E and F show a decreasing proportion in the control group membership as the profile distance increases, which is as expected since a larger distance indicates the recall is less similar to the original story. The differences evidenced in these plots motivate a modeling approach to quantify the different behaviors of the features.

The conditional density plots from Fig. 5 are consistent with the claim of siblings being ‘intermediate’ between controls and patients (e.g., Egan et al., 2001, 2003). This ordering of the categories suggests modeling groups (patients, siblings, controls) as an ordinal categorical variable and specifically supports a model based on a proportional odds logistic regression (Agresti, 2007; Venables & Ripley, 2002). Equation (2) provides a concrete example of a proportional odds model with human recall as an explanatory variable where log odds (the logit of the cumulative probability) of being in group j or below for individual i is a function of the group intercept a_j and parameters for recall-time t and the human score x_i on the Story 1 recall (Interested readers are referred to the references for details of estimation of this class of models.).

$$\mathrm{Logit}[P(Y_{it} <= j)] = a_j + b_t + b_{h1} \times x_i + e_i \tag{2}$$

Notice that this model can predict the group probabilities for each of the groups as differences between the appropriate cumulative probabilities. It follows that to predict the group with the best fit, requires selecting the group with the highest probability at these values of the parameters. Referring back to Fig. 5, we should not expect to predict members of the sibling group using a single variable, since for all of the variables, there is no value where sibling is the group with the largest proportion (which is the group that a prediction algorithm will choose). It is possible that a

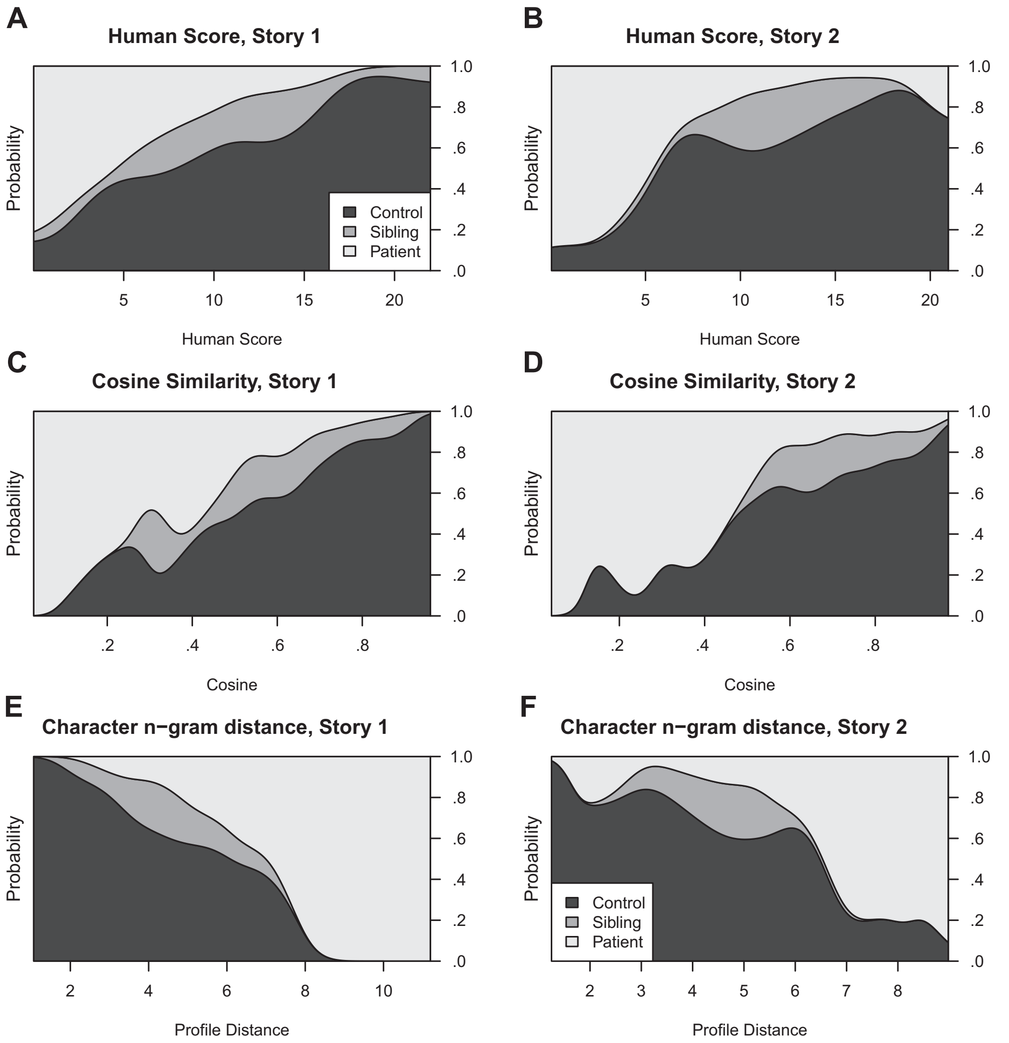

Fig. 5 – Panels A–F. Conditional density plots of probability of group membership based on values of features derived from recalls. Each row shows data from one of the features: human scores, LSA cosine similarity and character *n*-gram profile distance, while the columns show data from Story 1 and Story 2 respectively. Shading differentiates diagnosis.

combination of variables will predict some siblings, and thus models with multiple explanatory variables are explored. For the sibling group, what will be of interest from the modeling is the location of the region where siblings are relatively more probable and how that reflects the sibling regions shown in Fig. 5. As with the previous analysis, the model contains a factor for recall-time with levels of immediate, 30 min and 24 h recall, with immediate recall being the baseline level.

The approach examines each feature individually to see if models containing both stories improve on models with only a single story. We then examine models with multiple explanatory features and compare them to the single feature models. The initial model predicts group based on the human score for Story 1 and recall-time. The coefficients, standard errors, *t*-value and profiled confidence intervals for this model are presented in the upper panel of Table 7.

The estimates containing a vertical bar are the intercepts for the group boundaries, and are generally only of interest in computing the group probabilities. Since zero is not contained in the 95% confidence interval for any of the non-intercept parameter estimates, we conclude that the parameters are statistically significant. Fig. 6 illustrates the effects of this

Table 7 – Upper Panel: Modeling group on human score and recall-time, Story 1 only. Lower Panel: Modeling group on human score and recall-time, Stories 1 and 2.

	Estimate	Std. err.	t Value	2.5%	97.5%
Story 1 only					
Human Story 1	−.24	.03	−8.67	−.30	−.19
30 min	−.63	.29	−2.13	−1.21	−.06
24 h	−.79	.30	−2.61	−1.39	−.20
Control|Sibling	−2.65	.40	−6.59		
Sibling|Patient	−1.73	.39	−4.46		
Story 1 & 2					
Human Story 1	−.18	.04	−5.03	−.25	−.11
Human Story 2	−.11	.04	−3.01	−.18	−.04
30 min	−.59	.30	−1.98	−1.18	−.01
24 h	−.74	.30	−2.44	−1.35	−.15
Control|Sibling	−3.08	.43	−7.11		
Sibling|Patient	−2.13	.42	−5.14		

model's parameter estimates by plotting the group probability predictions for the three different recall-times as human score varies for Story 1.

This Figure provides a model-based view of the same data displayed in Fig. 5, but now represented with many fewer parameters, and the impact of recall-time explicitly represented. Color is used to indicate group and line type is used to indicate recall-time. Consider the three group curves for the immediate recall-time. For the predictions of the patient group, as the human score increases the probability of membership in the patient group decreases. The point where this curve intersects the increasing probability prediction curve for the control group occurs at a human score between 9 and 10, and divides the optimal prediction for a participant's group from patient with human scores in the range 0–9 to control for human scores 10 and above for the immediate recall-time. The plot clearly shows that the sibling group, while starting low and reaching its maximum probability of about .23 between human scores 10 and 11, is never more probable than the other two groups, the patient group below score 10 and then the control group for scores above 10.

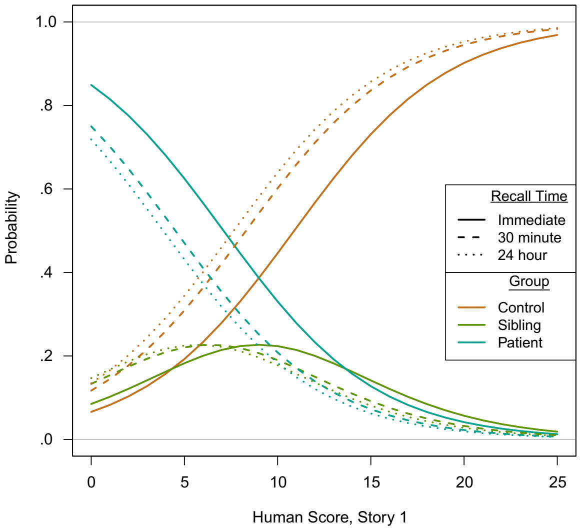

Fig. 6 – Model predictions by recall-time and group for human score for Story 1. Line type distinguishes recall-time and color distinguishes group. Human score ranges from 0 to 25.

With respect to recall-time, this model confirms the results of the linear mixed-effects models. In contrast to the immediate recall-time curve, the two longer recall curves are shifted to the left allowing similar probability of membership at lower scores. The vast majority of this shift occurs between immediate recall and 30 min for all three groups[10]. There is an interesting asymmetry to the plot. At the highest score levels the probability of membership in the control group is almost 1.0, while at the low end there is still some diversity in group membership probabilities. There is a probability of about .85 for the patient group, .08 for sibling and .07 for control. This may indicate an issue with the story so that the human score range could not extend quite low enough.

The next more complex model adds the human score for Story 2 as an explanatory variable. The coefficients for that model are shown in the lower panel of Table 7. All the coefficients are significant, which indicates that recalls on the second story add to the explanatory power of the model. It is noteworthy that the coefficients on the human scores for the two stories differ by over 50%. In this modeling format, it is not possible to determine if this is a statistically significant difference. In addition, these coefficients are odds ratios making interpretation slightly more complex, but it may merit further research to determine if the final WMS-R Logical Memory score, which is currently the sum of the human scores on each story is optimal in that a weighted score might provide a better measure. The likelihood ratio test (Table S5) is highly significant with $p = .0026$, and both AIC and BIC decrease with the more complex model indicating a strong preference for the model with explanatory human scores from both Stories 1 and 2.

The analyses are now repeated for the other two features, LSA cosine similarity and the character *n*-gram profile

[10] These same results can be derived from the parameter estimates, which are log odds ratios. The coefficient estimate on parameter Human Story 1 is −.24 (Table 7), which when exponentiated is .79, and can be interpreted that for a given recall-time, the higher the human score, the less are the odds to fall in the sibling or patient group in comparison to the control group. Similarly, the estimate on 30 min recall is −.63, which when exponentiated is .53. This indicates that the odds of being in the sibling or patient group is about 1/2 compared to the odds of being a control when moving from immediate recall to 30 min recall, which just indicates that at 30 min recall there is less difference in human scores between controls and the other two groups. In probability terms, refer to Fig. 6. For example, with a human score of 10, the probability of being in the control group is about .45 for immediate recall, but increases to .60 for 30 min recall. Converting from probability to odds ratio is [.45/(1 − .45)]/[.60/(1 − .60)] = .55 which is the exponentiated estimated parameter for 30 min recall, as expected (note there is rounding error due to only displaying two digits of accuracy).

Table 8 – Upper Panel: Modeling group on LSA cosine similarity and recall-time, Stories 1 and 2. Lower Panel: Modeling group on character *n*-gram profile distance and recall-time, Stories 1 and 2.

	Estimate	Std. err.	t Value	2.5%	97.5%
LSA cosine similarity					
Cosine Story 1	−5.49	.95	−5.77	−7.43	−3.69
Cosine Story 2	−3.96	.87	−4.56	−5.69	−2.28
30 min	−.65	.30	−2.18	−1.24	−.07
24 h	−.66	.30	−2.21	−1.25	−.08
Control|Sibling	−5.90	.76	−7.74		
Sibling|Patient	−4.95	.74	−6.68		
Character n-gram to Story					
n-gram Story 1	.61	.11	5.33	.39	.84
n-gram Story 2	.30	.13	2.38	.05	.55
30 min	−.75	.30	−2.50	−1.34	−.17
24 h	−.91	.31	−2.98	−1.52	−.32
Control|Sibling	4.23	.53	8.00		
Sibling|Patient	5.16	.55	9.30		

Table 9 – Confusion matrices. Upper Panel: Human scores model; Middle Panel: LSA Cosine similarity model; Lower Panel: *n*-gram Similarity model.

Human scores model			
Actual Group		Predicted Group	
	Control	Sibling	Patient
Control	199	0	21
Patient	43	0	37
Sibling	45	0	8
LSA Cosine similarity model			
Actual Group		Predicted Group	
	Control	Sibling	Patient
Control	206	0	14
Patient	36	0	44
Sibling	50	0	3
***n*-gram Similarity model**			
Actual Group		Predicted Group	
	Control	Sibling	Patient
Control Group	195	0	25
Patient	33	0	47
Sibling	51	0	2

distance as explanatory variables. Table 8 shows the coefficients for the models including both Story 1 and Story 2. For both features, all the coefficients are significant as are the likelihood ratio tests (see Table S6).

The counterpart plot for cosine similarity to Fig. 6 is shown in Fig. 7. The overall story is much the same, though the curves are more symmetric with respect to the range of cosine similarity. The leftward shifts for the 30 min and 24 h recall-times significantly separate those curves (as evidenced by their significant coefficient estimates) from the immediate recall, but for all groups these two curves are quite similar to each other. The trade-off between predicting the patient versus the control group occurs at a cosine similarity between .50 and .55.

At this point, we have established that for each feature using information from both stories yields a better model. The last step is to compare the predictive ability of each of these features and examine a model that combines the features. Predictions based on the same observations that a model is trained on yields biased (optimistic) results (see for instance Hastie, Tibshirani, & Friedman, 2009), therefore the predictive performance is evaluated using 10-fold cross validation. Breaking out observations by group, there were 220 for controls, 53 for siblings and 80 for patients, and thus any performance should be judged against just predicting the largest group namely, the control group, for all observations, which yields an accuracy of 62.3%. Assuming an equal penalty for any incorrect group prediction, this accuracy is the prediction performance to match or exceed.

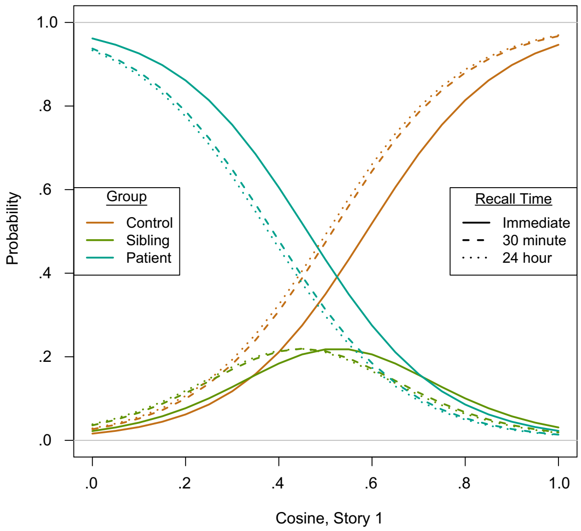

Fig. 7 – Model predictions by recall-time and group for cosine similarity measure for Story 1. Line type distinguishes recall-time and color distinguishes group. Cosines range from .0 to 1.0.

Table 10 – Potential variables for stepwise model selection. The last two variables were only available for the model including human scores.

Recall-time
Cosine Story 1
Cosine Story 2
n-gram Story 1
n-gram Story 2
n-gram English Story 1
n-gram English Story 2
Human Story 1
Human Story 2

Table 11 – Stepwise models. Upper Panel: model without human scores and Lower Panel: model with human scores.

Model without human scores	Estimate
Control|Sibling	.35
Sibling|Patient	1.34
30 min	−.73
24 h	−.84
n-gram Story 1	.45
Cosine Story 2	−4.38
n-gram English Story 1	.60
n-gram English Story 2	−.38

Model with human scores	Estimate
Control|Sibling	−2.84
Sibling|Patient	−1.84
30 min	−.66
24 h	−.79
Human Story 1	−.14
Cosine Story 2	−4.15
n-gram English Story 1	.66
n-gram English Story 2	−.38

Table 9 shows the results of running cross-validated models for human scores, LSA cosine similarity and the *n*-gram measure. The upper panel contains the confusion matrix from the model using human scores as explanatory variables and we see for instance from the first row that of the 220 controls, 199 are correctly predicted as controls, 0 are predicted as siblings and 21 are incorrectly predicted as patients. The middle panel shows the confusion matrix from the model using LSA cosine similarities as the explanatory variables, where now the first row shows that now 206 of the controls are correctly predicted and only 14 are incorrectly predicted as patients and the lower panel shows the predictions from the *n*-gram model.

The model with human scores as the explanatory variable has an accuracy of 66.9%, which is above the baseline guessing accuracy of 62.3%, while the semantic model prediction accuracy was 70.8%. The *n*-gram model prediction accuracy was 68.5%, but tended to classify patients more accurately than either humans or the semantic model, though also classified many more controls as patients than the other two models. As noted earlier in discussing the conditional density plots (see Fig. 5), we did not expect (and so it turned out), that no participant was predicted in the sibling group from either model. This confirms that these simple models are not able to distinguish the siblings group. To improve the quality of the estimates, the cross-validation runs were repeated and the estimates of the accuracy were averaged for all the models. Before discussing improving estimates, we first discuss issues involved in the construction of models combining the features.

Table 12 – Prediction performance of models with varying feature sets.

Model features	Mean accuracy	Std. dev.
n-grams to Story	66.82	.61
n-grams to English	67.93	.59
Human scores	68.21	.65
Cosines	70.11	.64
Combined	70.31	.69
Combined + human	70.43	.69

To combine these features into a single model we used a stepwise model selection based on AIC, since that fits in well with the use of cross-validation prediction to validate the choices. Though there are issues with stepwise model selection in stability of feature choice with collinear features and with bias in the standard errors, it is still valid to make inferences over the model predictions.

Two stepwise models were generated based on slightly different initial variable sets. The first did not include the human scores as potential features using only the automated features, and the second included human scores as potential features. The performance of these two models allows judging if there is any additional predictive power in predicting group remaining in the human scores after accounting for the semantic and character *n*-gram sequential features. We use the stepAIC function from the MASS package (Venables & Ripley, 2002) of R to perform the stepwise model selection. Table 10 shows the starting variable set, and Table 11 shows the models (the variables and their coefficient estimates) resulting from the stepwise selection for the variable set without human scores, and including human scores.

The results indicated that the model that included human scores amongst its potential predictors swapped the character *n*-gram recall to Story 1 with the human score on Story 1, which is another indication of how closely the character *n*-gram modeled human scores. The model with human score had a smaller AIC at 542.87 in contrast to 547.501 for the previous model. We next performed a 10-fold cross-validation prediction repeated 1000 times for each of the models. The mean exact agreement and standard deviation are shown in Table 12.

Despite the high correlation to human scores, the character *n*-gram to the story predicts least well with exact agreement of 66.82%. Next best is the character *n*-gram to the English profile at 67.93%. Next are the human scores at 68.21%. The LSA semantic measures are at 70.11%. Finally, the stepwise model that includes both types of language features is slightly better at 70.31% and including a human score brings the best model up to 70.43%, so there is a very small part of the human variance not captured by the other automated measures, though the model based on just the semantic and sequential features does quite well. Given the large number of repetitions, all the differences between these models are statistically significant. For instance the *t*-test contrasting the combined model and combined model plus human scores is highly significant, $t(999) = 5.90$, $p < .001$, as is the difference between the character *n*-gram to recall versus the character *n*-gram to English profile with $t(999) = -49.92$, $p < .001$.

4. Discussion

Consistent with previous reports (e.g., Egan et al., 2003; Skelley et al., 2008), patients with schizophrenia performed significantly more poorly than healthy controls on a widely-used and well-respected measure of verbal episodic memory function, with siblings performing intermediary between the two groups. The largest drop in performance across all three groups was seen from immediate to short-delay (30 min) recall, with a much smaller decline from 30 min to 24 h. This is consistent with theories of memory consolidation, and in the main replicates previous findings (e.g., Skelley et al., 2008). However, there are differences in that the study of Skelley et al. (2008) found siblings' performance to be poorer than controls, whereas we found a significant difference only for Story 1 and the difference was not strong enough to survive in the summed score. In terms of savings scores, we also found a slightly different pattern to Skelley et al. (2008) at both short and long delays. Although we also (not surprisingly) found patients' performance was poorer than controls at short delay, we also found they recalled significantly less than siblings [which Skelley et al. (2008) did not, although they did find siblings' performance poorer than controls at short delay]. Furthermore, we did not uncover any differences at long delay, whereas Skelley et al. (2008) reported worse performance in patients relative to controls at long delay. Interestingly, both our and Skelley et al.'s (2008) patient group performed remarkably well in terms of long-delay savings [our study – 89%; Skelley et al. (2008) – 87%] compared with short delay [our study – 58%; Skelley et al. (2008) – 66%]. Demographic differences between the samples do not seem to be large enough to account for observed differences (although we note that our sample is considerably smaller and younger), and thus it is possible that the observed differences across studies reflect power issues in our sample (as it is considerably smaller). It is also possible that some of these differences (specifically the saving score differences) are an artifact of not dividing both 30 min and 24 h by immediate recall. Given our results, there are large individual differences. One way to compensate is to divide through by the immediate score and hope that one is removing the individual component. The alternative approach is the one we have adopted in this paper; namely of using a linear mixed model.

Beyond these differences across studies, our modeling revealed robust correlations between the human raters and both the language sequential features and LSA-based semantic similarity features. This suggests that the automated computational approach is both valid and reliable as a complimentary scoring method to humans on this task. Despite a high correlation of human rating scores with our measures of sequential features, we conclude that LSA performed better (than humans or syntax alone) at detecting diagnostic group differences. While the human inter-rater reliability values were marginally stronger than the LSA-human inter-rater reliability values, they were nonetheless uniformly high, and it is important to keep in mind that the two methods (human and machine) were employing different scoring strategies to achieve the same end. Human raters in essence matched specific words, whereas the LSA cosine matched passages based on overall meaning, without regard to the presence of key words or word order. A rating task like matching words, because it is so constrained, can drive high levels of consistency in human raters even if that consistency is not always measuring the construct of interest. More important to note, while the computational model performed qualitatively similarly to the human raters, it was in fact significantly better able to predict group membership across three time points (i.e., immediate recall, 30 min later and 24 h later) based on participants' test scores. This demonstrates that LSA variables can be used *interchangeably* with human ratings, and may well provide both more accurate and detailed information. Further research is needed to examine other aspects of recall (e.g., omissions and tangentiality) to determine how well LSA variables can capture these variables that are not recognized under current scoring rubrics. While logistic regression has a number of useful features, in future work for predicting group, more sophisticated machine learning classification techniques, such as support vector machines (Cortes & Vapnik, 1995) can also be considered. In addition, we note that in this paper we have only used the sequential features as scalars. Just as with the semantic measures, one could consider using a *k*-near measure (namely find the nearest *n*-gram neighbors and select the group as the majority group among the *k*-near set).

In this paper, we have employed recalls from a widely used prose recall test to evaluate a framework that incorporated analyses of both semantic and sequential language features that may be implicated in verbal recall. The framework permitted the exploration of the usefulness of an automated scoring methodology that has the potential to provide similar or better scoring metrics to that of human raters, as well as a more detailed characterization of recall performance over time. The strong predictions of the models in this framework indicate that these language features can be closely implicated in differences in language from people in different clinical states. Concerning this latter issue, we suggest that the framework we have presented may help in the much needed enterprise of defining the behavioral phenotype that may relate more directly to underlying neurobiology and how genes effect neural systems and behavior. Indeed, a core premise underlying the current computational approach to prose recall is that a more fine-grained framework with which to parse the components of prose recall – in this case especially its language sequences and semantic parts – will be of use in unraveling some of the hallmark deficits of episodic verbal memory in schizophrenia, and thereby contribute to a greater understanding of the underlying cognitive and neural mechanisms.

A complementary approach exploiting computational models can provide additional insights into the sources of language disruption. Hoffman et al. (2010) utilized models based on neural networks and in concordance with the results seen here, found evidence of memory consolidation failures attributable for language patterns found in schizophrenia. The promise of these computational models, especially with the recent advances in developing emergent features from "deep" networks (Hinton, Osindero, & Teh, 2006), may help uncover new, potentially more diagnostic phenotypes. Although interesting, a discussion of this area is beyond the scope of this paper, but serves to emphasize the complex and intertwined nature of episodic memory and the semantic and sequential aspects of language. Put differently, the framework that we have presented illustrates a way in which the actual words that are uttered may be used as a critical tool with which to explore the neurocognitive mechanisms and systems underlying prose recall.

Acknowledgments

We gratefully acknowledge Daniel R. Weinberger M.D., for making the data available, which were collected via the Clinical Brain Disorders (NIMH) Schizophrenia Sibling Study (D.R. Weinberger, PI). This research was supported by the Intramural Research Program of the National Institute of Mental Health, National Institutes of Health, USA. We are also thankful for assistance from Julia Longenecker M. Phil. and Krista Wisner B.A. Brita Elvevåg was funded by the Northern Norwegian Regional Health Authority, Helse Nord RHF, and Peter Foltz and Mark Rosenstein are employees of Pearson, the publisher of the Wechsler test utilized in the paper, which partially supported their work as part of their employment. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix. Family random effect

Both human score and cosine similarity responses were modeled using group and recall-time as fixed effect explanatory variables within a linear mixed effects framework. The impact of adding a random effect for family to a baseline model with an existing individual random effect was investigated to determine if the more complex model generated a more predictive model. The hypothesis is that some of the ability in the recall task coheres within families (e.g., Egan et al., 2001, 2003), and it would improve the quality of the estimates to correctly account for the variance between individual and family effects if that proves necessary. Unfortunately there are issues with this data set that negatively impact the ability to discriminate family effects. The control group, which constitutes the largest group, has no siblings in the study and families with more than one member are fairly rare among the other two groups including 17 patients without siblings and five siblings without patients. There are 122 unique individuals comprising 109 families. Table A1 shows the family IDs (FID) (randomly assigned for this analysis) and the counts for the 11 families that are represented by more than one individual. Two families are composed of two siblings and one patient, and the rest with one sibling and one patient. This shortage of multi-member families provides the statistical machinery with less information on family performance so likely will have increased difficult in allocating variance between individual effects and family effects.

Table A1 – Family member counts for families with more than one member.

FID	C	S	P
8A3	0	2	1
8A4	0	1	1
8A8	0	2	1
8A9	0	1	1
8B3	0	1	1
8B9	0	1	1
8C3	0	1	1
8D5	0	1	1
8D7	0	1	1
8D8	0	1	1
8F2	0	1	1

Using notation described in Methods section, we will compare the models specified in equation (A1) with only an individual random effect and (A2), which adds a family random effect, z_f. A likelihood ratio test as well as AIC and BIC will guide the comparison of the models. We will separately test the response Y with human scores and cosine similarities with fixed effects for group (g) and recall-time (t), and a random effect for individual z_i. The fixed effects for equation (A1) for both Stories 1 and 2 are presented in Table A2. See the description for Table 5 for more details on how to interpret the fixed effects. Remodeling adding a family random effect yields nearly identical fixed effects, which are not repeated here.

$$Y_{it} = b_0 + b_g + b_t + z_i + e_i \tag{A1}$$

$$Y_{it} = b_0 + b_g + b_t + z_i + z_f + e_i \tag{A2}$$

Table A2 – Fixed effects modeling human scores for Stories 1 and 2 without a family random effect.

	Estimate	Std. err.	t Value	2.5%	97.5%
Story 1					
Fixed effects:					
Intercept	14.92	.49	30.53	13.96	15.87
Sibling	−1.65	1.09	−1.52	−3.77	.47
Patient	−6.45	.92	−7.03	−8.24	−4.66
30 min	−2.53	.20	−12.89	−2.92	−2.15
24 h	−2.88	.20	−14.45	−3.28	−2.49
Story 2					
Fixed effects:					
Intercept	13.48	.47	28.77	12.56	14.39
Sibling	−.43	1.04	−.41	−2.46	1.60
Patient	−5.70	.88	−6.48	−7.42	−3.98
30 min	−1.16	.18	−6.29	−1.53	−.80
24 h	−1.30	.19	−6.89	−1.66	−.93

Table A3 shows the random effects for the two models with and without a family effect for Story 1. Between the two models, the residual variance remains the same, so in the family model the variance explained by the random effects is just reallocated between individual and family. The 95% confidence interval for the standard deviation for the family random effect includes zero (represented as a lower bound of undefined denoted by NA) so it is not significant (Since standard deviation cannot assume values less than zero, the lower bound becomes undefined, hence the NA). Story 2 was also modeled without and with family, and the fixed effects were nearly identical between models. The 95% confidence interval for the standard deviation of the family effect includes zero, so the Story 2 coefficients are not presented here.

Table A3 – Random effects modeling human scores for Story 1 with and without a family random effect.

Groups	Name	Var.	Std. dev.	2.5%	97.5%
ID only					
Random effects:					
ID	Intercept	16.38	4.05	3.52	4.58
Residual		2.29	1.51	1.38	1.65
Number of obs: 353; ID: 122					
ID + family					
Random effects:					
ID	Intercept	12.72	3.57	2.32	4.55
Family	Intercept	3.70	1.92	NA	3.54
Residual		2.29	1.51	1.38	1.65
Number of obs: 353; ID: 122; family: 109					

Table A4 shows the results of comparing the two models. The AIC increased for Story 1 and was identical for Story 2 while the BIC increased for both of the more complex models. In neither case was the likelihood ratio test significant, so we conclude the more complex model with a family random effect is not more predictive.

Table A4 – Likelihood ratio tests for Story 1 and Story 2 modeling human scores with and without family random effect.

	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	Pr (>Chisq)
Story 1								
ID only	7	1676.7	1703.8	−831.35	1662.7			
ID + family	8	1678.3	1709.2	−831.15	1662.3	.40	1	.53
Story 2								
ID only	7	1639.4	1666.4	−812.68	1625.4			
ID + family	8	1639.4	1670.3	−811.71	1623.4	1.94	1	.16

The analysis was repeated, but this time using cosine similarity as the response variable. Table A5 shows the estimates for Story 1. In this case, there is a small drop in variance of the residuals between the models, though the estimate of the standard deviation of the family effect, as before, contains zero and is so not significant. The estimates of the fixed effects do not change much between the models and are not presented here.

Table A5 – Random effects modeling cosine for Story 1 with and without a family random effect.

Groups	Name	Var.	Std. dev.	2.5%	97.5%
ID only					
Random effects:					
ID	Intercept	.0073	.086	.11	.14
Residual		.0073	.086	.062	.074
Number of obs: 353; ID: 122					
ID + family					
Random effects:					
ID	Intercept	.0055	.074	.052	.13
Family	Intercept	.0055	.074	NA	.12
Residual		.0055	.074	.062	.074
Number of obs: 353; ID: 122; family: 109					

Table A6 shows the results of the likelihood ratio test between the two models. AIC is smaller for the more complex model, but BIC is larger. The p-value is .077. While not significant, (Pinheiro & Bates, 2000) cited in (Bates to be published) have shown that probabilities generated at the boundary of the legal values, as is the case here for the standard deviation, are conservative, by up to a factor of 2. Stepping back and computing the 90% confidence interval (from 5% to 95%) for the family standard deviation is (.024, .12), which does not include zero so indicates significance at the 90% level.

The analysis was repeated for Story 2. There was a small drop in residual variation adding the family random effect, but the standard deviation was not significant. Table A6 shows that the AIC was the same for the models, while the BIC increased in the more complex model and with a $p = .15$, the likelihood ratio was not significant.

There is a small hint that there might be a family effect, since the standard deviation for the cosine response in Story 1 was nearly significant. We do not want to overstate the importance of this result since the three other cases were clearly not significant. Given the scarcity of families in this data set it does suggest that further investigation of a family effect may be warranted. This also may suggest that the cosine similarity is capturing some part of the recall phenomena that the human scores are not.

Table A6 – Likelihood ratio tests for Story 1 and Story 2 modeling cosine with and without family random effect.

	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	Pr (>Chisq)
Story 1								
ID only	7	−601.34	−574.28	307.67	−615.34			
ID + family	8	−602.48	−571.55	309.24	−618.48	3.13	1	.077
Story 2								
ID only	7	−548.50	−521.43	281.25	−562.50			
ID + family	8	−548.54	−517.61	282.27	−564.54	2.04	1	.15

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cortex.2014.01.021.

REFERENCES

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Aleman, A., Hijman, R., de Haan, E. H. F., & Kahn, R. S. (1999). Memory impairment in schizophrenia: a meta-analysis. *American Journal of Psychiatry, 156*, 1358–1366.

Armstrong-Warwick, S., Thompson, H. S., McKelvie, D., & Petitpierre, D. (1994). Data in your language: the ECI multilingual corpus 1. In *Proceedings of the International workshop on sharable natural language resources. Nara, Japan* (pp. 97–106).

Baddeley, A., & Wilson, B. A. (2002). Prose recall and amnesia: implications for the structure of working memory. *Neuropsychologia, 40*, 1737–1743.

Baitz, H. A., Thornton, A. E., Procyshyn, R. M., Smith, G. N., MacEwan, G. W., Kopala, L. C., et al. (2012). Antipsychotic medications: linking receptor antagonism to neuropsychological functioning in first episode psychosis. *Journal of the International Neuropsychological Society, 18*(4), 717–727.

Barch, D. M. (2005). The cognitive neuroscience of schizophrenia. *Annual Review of Clinical Psychology, 1*, 321–353.

Bates, D. M. (draft 2010). lme4: Mixed-effects modeling with R. Unpublished book draft. Retrieved from http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 0.999902345-0 http://lme4.r-forge.r-project.org/.

Brébion, G., Bressan, R. A., Amador, X., Malaspina, D., & Gorman, J. M. (2004). Medications and verbal memory impairment in schizophrenia: the role of anticholinergic drugs. *Psychological Medicine, 34*(2), 369–374.

Cabana, Á., Valle-Lisboa, J. C., Elvevåg, B., & Mizraji, E. (2011). Detecting order-disorder transitions in discourse: implications for schizophrenia. *Schizophrenia Research, 131*, 157–164.

Cavnar, W. B., & Trenkle, J. M. (1994). N-Gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (pp. 161–175).

Cirillo, M. A., & Seidman, L. J. (2003). Verbal declarative memory dysfunction in schizophrenia: from clinical assessment to genetics and brain mechanisms. *Neuropsychology Review, 13*, 43–77.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science, 41*(6), 391–407.

Dunn, J. C., Almeida, O. P., Barclay, L., Waterreus, A., & Flicker, L. (2002). Latent semantic analysis: a new method to measure prose recall. *Journal of Clinical and Experimental Neuropsychology, 24*(1), 26–35.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology,* (H. Ruger & C. Bussenius, Trans.). New York, NY: Teachers College (Original work published 1885).

Egan, M. F., Goldberg, T. E., Gscheidle, T., Weirich, M., Bigelow, L. B., & Weinberger, D. R. (2000). Relative risk of attention deficits in siblings of patients with schizophrenia. *American Journal of Psychiatry, 157*, 1309–1316.

Egan, M. F., Goldberg, T. E., Gscheidle, T., Weirich, M., Rawlings, R., Hyde, T. M., et al. (2001). Relative risk for cognitive impairments in siblings of patients with schizophrenia. *Biological Psychiatry, 50*(2), 98–107.

Egan, M. F., Kojima, M., Callicott, J. H., Goldberg, T. E., Kolachana, B. S., Bertolino, A., et al. (2003). The BDNF val66met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell, 112*, 257–269.

Elvevåg, B., Foltz, P. W., Rosenstein, M., & DeLisi, L. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics, 23*, 270–284.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Research, 93*, 304–316.

Elvevåg, B., & Goldberg, T. E. (2000). Cognitive impairment in schizophrenia is the core of the disorder. *Critical Reviews in Neurobiology, 14*, 1–21.

First, M., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997). *User's guide for the structured clinical interview for DSM-IV axis I disorders – Clinician version (SCID-CV)*. Washington, DC: American Psychiatric Press.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: applications to educational technology. In B. Collis, & R. Oliver (Eds.), *Proceedings of EDMedia '99* (pp. 939–944). Charlottesville, VA: Association of Computing in Education.

Goldberg, T. E., Torrey, E. F., Gold, J. M., Bigelow, L. B., Ragland, R. D., Taylor, E., et al. (1995). Genetic risk of neuropsychological impairment in schizophrenia: a study of monozygotic twins discordant and concordant for the disorder. *Schizophrenia Research, 17*, 77–84.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*(7), 1527–1554.

Ho, B. C., Milev, P., O'Leary, D. S., Librant, A., Andreasen, N. C., & Wassink, T. H. (2006). Cognitive and magnetic resonance imaging brain morphometric correlates of brain-derived neurotrophic factor val66met gene polymorphism in patients with schizophrenia and healthy volunteers. *Archives of General Psychiatry, 63*, 731–740.

Hoffman, R., Grasemann, U., Gueorguieva, R., Quinlan, D., Lane, D., & Miikkulainen, R. (2010). Using computational patients to evaluate illness mechanisms in schizophrenia. *Biological Psychiatry, 69*(10), 997–1005.

Hofmann, H., & Theus, M. (2005). *Interactive graphics for visualizing conditional distributions*. Unpublished Manuscript. (cited in R Core Team (2012)).

Hornik, K., Rauch, J., Buchta, C., & Feinerer, I. (2012). *Textcat: N-Gram based text categorization*. R package version 0.1-1 http://CRAN.R-project.org/package=textcat.

Jastak, S., & Wilkinson, G. S. (1984). *The wide range achievement test: Revised administration manual* (rev ed.). Wilmington, DE: Jastak Associates, Inc.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Kalkstein, S., Hurford, I., & Gur, R. C. (2010). Neurocognition in schizophrenia. *Current Topics in Behavioral Neuroscience, 4*, 373–390.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79–86.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Lautenschlager, N. T., Dunn, J. C., Bonney, K., Flicker, L., & Almeida, O. P. (2006). Latent semantic analysis: an improved method to measure cognitive performance in subjects of non-English speaking background. *Journal of Clinical and Experimental Neuropsychology, 28*(8), 1381–1387.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.

Lim, K. O., Ardekani, B. A., Nierenberg, J., Butler, P. D., Javitt, D. C., & Hoptman, M. J. (2006). Voxelwise correlational analyses of white matter integrity in multiple cognitive domains in schizophrenia. *American Journal of Psychiatry, 163*(11), 2008–2010.

Longenecker, J., Genderson, J., Dickinson, D., Malley, J., Elvevåg, B., Weinberger, D. R., et al. (2010). Where have all the women gone? Participant gender in epidemiological and non-epidemiological research of schizophrenia. *Schizophrenia Research, 119*, 240–245.

Longenecker, J., Kohn, P., Liu, S., Zoltick, B., Weinberger, D. R., & Elvevåg, B. (2010). Data-driven methodology illustrating mechanisms underlying word list recall: applications to clinical research. *Neuropsychology, 24*, 625–636.

Matsui, M., Sumiyoshi, T., Abe, R., Kato, K., Yuuki, H., & Kurachi, M. (2007). Impairment of story memory organization in patients with schizophrenia. *Psychiatry and Clinical Neurosciences, 61*, 437–440.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society Series B, 42*, 109–142.

Missar, C. D., Gold, J. M., & Goldberg, T. E. (1994). WAIS-R short forms in chronic schizophrenia. *Schizophrenia Research, 12*, 247–250.

Mori, K., Nagao, M., Yamashita, H., Morinobu, S., & Yamawaki, S. (2004). Effect of switching to atypical antipsychotics on memory in patients with chronic schizophrenia. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 28*(4), 659–665.

Munro Cullum, C., Butters, N., Tröster, A. I., & Salmon, D. P. (1990). Normal aging and forgetting rates on the Wechsler Memory Scale-Revised. *Archives of Clinical Neuropsychology, 5*(1), 23–30.

O'Driscoll, G. A., Florencio, P. S., Gagnon, D., Wolff, A. V., Benkelfat, C., Mikula, L., et al. (2001). Amygdala-hippocampal volume and verbal memory in first-degree relatives of schizophrenic patients. *Psychiatry Research, 107*(2), 75–85.

Pinheiro, J., & Bates, D. (2000). *Mixed-Effects models in S and S-PLUS*. New York: Springer.

Rabin, L., Barr, W., & Burton, L. (2005). Assessment practices of North American Clinical Psychologists: a survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*(1), 33–65.

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0. http://www.R-project.org/.

Robinson, E. S., & Heron, W. T. (1922). Results of variations in length of memorized materials. *Journal of Experimental Psychology, 5*(6), 428–447.

Russell, E. W. (1988). Renorming Russell's version of the Wechsler memory scale. *Journal of Clinical and Experimental Neuropsychology, 10*(2), 235–249.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.

Skelley, S. L., Goldberg, T. E., Egan, M. F., Weinberger, D. R., & Gold, J. M. (2008). Verbal and visual memory: characterizing the clinical and intermediate phenotype in schizophrenia. *Schizophrenia Research, 105*, 78–85.

Toulopoulou, T., Rabe-Hesketh, S., King, H., Murray, R. M., & Morris, R. G. (2003). Episodic memory in schizophrenic patients and their relatives. *Schizophrenia Research, 63*(3), 261–271.

Tröster, A. I., Butters, N., Salmon, D. P., Cullum, C. M., Jacobs, D., Brandt, J., et al. (1993). The diagnostic utility of savings scores: differentiating Alzheimer's and Huntington's diseases with the logical memory and visual reproduction tests. *Journal of Clinical and Experimental Neuropsychology, 15*(5), 773–788.

Vassos, E., Bramon, E., Picchioni, M., Walshe, M., Filbey, F. M., Kravariti, E., et al. (2010). Evidence of association of KIBRA genotype with episodic memory in families of psychotic patients and controls. *Journal of Psychiatric Research, 44*, 795–798.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Wechsler, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology, 19*, 87–95.

Wechsler, D. (1987). *Wechsler Memory Scale – Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Memory Scale – Third Edition, WMS-III: Administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2009). *Wechsler Memory Scale—Fourth Edition (WMS–IV) technical and interpretive manual*. San Antonio, TX: Pearson.

Weickert, T. W., Goldberg, T. E., Gold, J. M., Bigelow, L. B., Egan, M. F., & Weinberger, D. R. (2000). Cognitive impairments in patients with schizophrenia displaying preserved and compromised intellect. *Archives of General Psychiatry, 57*, 907–913.

Wiens, A. N., Bryan, J. E., & Crossen, J. R. (1993). Estimating WAIS-R FSIQ from the national adult reading test-revised in normal subjects. *The Clinical Neuropsychologist, 7*, 70–84.

Zeno, S., Ivens, S., Millard, R., & Duvvuir, R. (1995). *The educator's word frequency guide*. Touchstone Applied Science Associates (TASA), Inc.

Zipf, G. K. (1935). *The psychobiology of language*. Boston, MA: Houghton-Mifflin.