Psychiatry Research 294 (2020) 113494

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

A Dynamic Method, Analysis, and Model of Short-Term Memory for Serial Order with Clinical Applications

Sophia Cheng[a], Alex S. Cohen[b], Terje B. Holmlund[c], Peter W. Foltz[d], Jian Cheng[e], Jared C. Bernstein[e], Elizabeth P. Rosenfeld[e], Brita Elvevåg[c,f,*]

[a] Henry M. Gunn High School, Palo Alto, California 94306, USA
[b] Department of Psychology, Louisiana State University, Baton Rouge, Louisiana, USA
[c] Department of Clinical Medicine, University of Tromsø - The Arctic University of Norway, Tromsø, Norway
[d] Institute of Cognitive Science, University of Colorado Boulder, Colorado, USA
[e] Analytic Measures Inc., Palo Alto, California, USA
[f] Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway

ARTICLE INFO

Keywords:
Temporal dynamics
Memory retrieval time
Smart-device

ABSTRACT

This study examined the robustness of a traditional memory task when moved out of controlled traditional settings. A letter recall task was designed to be self-administered via a smart-device which assessed recall by participants' writing their responses on the device. This enabled collection of both the letter recalled and the timing of this recall such that the temporal dynamics could be examined. Participants were patients with mental illness ($n=71$) and healthy volunteers ($n=103$). Temporal dynamics were examined using a new mechanism that measured memory retrieval time precisely. Data were analyzed for accuracy, time and their relationships. The classic memory phenomena and associated effects were replicated. In terms of temporal dynamics, this is the first demonstration of primacy and recency effects in time domain variables, as well as phonological similarity effects as evident by the inverted U-shaped curves in time. The speed of short-term memory processes affects accuracy, error types and timing. The robustness of these memory effects and new approach to temporal dynamics suggest this framework may be suitable for clinical applications, notably for the long-term monitoring of cognition in patients with mental illness.

1. Introduction

Memory is at the core of our daily lives. Short-term memory (STM) is needed for a wide variety of behaviors in humans, and understanding the mechanism behind poorer recall promises to improve our understanding of serious mental illness. STM includes the processing, storage and retrieval of sequential information which is critical for remembering a series of numbers such as when glancing at one's phone and looking away to confirm the license plate number of a vehicle, but also for more complex processes such as the planning of utterances in conversation and the associated flow of thoughts. Memory displays reliable and robust patterns, including the primacy and recency effects (e.g., Deese, 1957; Jahnke, 1965), wherein items presented early and late in a sequence are recollected better than other items. The phonological similarity effect (Conrad & Hull, 1964) is another core finding in STM research, wherein items that are phonologically non-confusable (e.g., M,J,H,Z,Q,Y) are recollected far better than items that are phonologically confusable (e.g., P,T,D,G,V,C).

The current study focused on the processing, storage and retrieval of sequential information in STM, which has been a topic of intense investigation for more than a century (e.g., Botvinick & Plaut, 2006; Ebbinghaus, 1885; Nipher, 1876; Thomas et al., 2003). Many theories have been proposed to account for recall patterns (e.g., Botvinick & Plaut, 2006; Brown et al., 2007; Farrell, 2006; Farrell & Lewandowsky, 2002; Henson, 1998; Page & Norris, 1998), and a variety of mechanisms have been proposed to account for information loss, including interference and time (e.g., Cowan & AuBuchon, 2008; Jonides et al., 2008; Lewandowsky et al., 2004; Oberauer & Lewandowsky, 2008). Research on the temporal dynamics of serial recall (e.g., Farrell & Lewandowsky, 2004; Kahana & Jacobs, 2000; Maybery et al., 2002; Oberauer, 2003; Thomas et al., 2003) has examined inter-response times that may serve as a rough proxy for the subsequent retrieval times. Surprisingly, the mechanisms underlying STM and the role of time in the performance pattern are still hotly

* Corresponding author. Department of Clinical Medicine, University Hospital of North Norway-Åsgård, Postbox 6124, 9291 Tromsø, Norway
E-mail address: brita@elvevaag.net (B. Elvevåg).

https://doi.org/10.1016/j.psychres.2020.113494
Received 15 June 2020; Accepted 29 September 2020
Available online 06 October 2020

debated, even at the most fundamental levels, but there has been a notable absence of research specifically examining the role of temporal dynamics in serial recall performance. The current study sought to examine these temporal dynamics in order to clarify the relationship between recall timing and the subsequent accuracy patterns. Specifically, patients with mental illness were included in this study to establish 'proof of concept' in the target population and to improve generalizability. As such this study was not designed to make specific inferences about how mental illness impacts STM but rather to assess the viability of technology for the collection, automatic transcription and analysis of STM data in various populations including mental illness.

2. Method

2.1. Participants

The data were collected through a study approved by the Louisiana State University (LSU) Institutional Review Board (#3618), and all participants provided their written informed consent. One group of participants comprised inpatients with mental illness who were in a treatment program for substance abuse disorder and who were paid for their participation ($5 per completed session). The other group of participants comprised undergraduate students recruited from a non-patient setting (henceforth termed 'nonpatients'). By using smart-devices, the study was conducted outside of a traditional laboratory setting, although patients participated in the presence of research assistants or clinical personnel. All individuals were asked to take the recall task at least once, then later up to three times, several days after the previous one. In total, 106 nonpatient participants and 93 patients took part. Out of the 106 nonpatients, 68 completed the task once only (i.e., one session), whereas 37 did two sessions, and one completed three sessions. The patients' distribution was more uniform, with 24 completing one session only, and 37 who did two sessions, and 32 who took part in all three sessions. Among these, 22 patients and 3 nonpatient participants produced clearly non-compliant data and thus their data were excluded from the analysis. This left 103 nonpatient and 71 inpatients whose data were analyzed here.

2.2. Serial letter recall

The serial letter recall task is part of the *delta* Mental State Examination or *d*MSE (Analytic Measures Inc., 2018), a research tool administered via an iOS app for the purpose of remotely monitoring the mental and cognitive health of participants (see Cheng et al., 2018; Chandler et al., 2019, 2020; Cohen et al., 2019; Holmlund et al., 2019a, 2019b, 2020). The *d*MSE employed neuropsychological tasks that were similar to the tasks in standard, widely-used neuropsychological tests. Participants were asked to take the tests regularly, but no more than once per day. In the *d*MSE sessions, participants were provided with short engaging tasks that required them to watch, listen, speak, and do screen touches in order to interact with the smart-device. An entire test session took roughly 11 minutes, although the serial recall part of the session took less than 2 minutes. This serial recall task occurred once every four times the app was used (on average once every eight days), and consisted of four recall tasks: two series with six phonologically confusable letters (e.g., P,T,D,G,V,C) and two series of non-confusable letters (e.g., M,J,H,Z,Q,Y). Each set was randomly chosen from a set of 16 predetermined sequences, which were a balanced list of different letters, all at equal difficulty levels. Six letters were chosen because six items are presumed to be at the threshold for people with poorer working memories, but still not too easy for those with good working memories, hence reducing the risk of ceiling and floor effects (for details, see Henson, 1998; Elvevåg et al., 2001; but unlike these studies which employed at least 20 trials of each list, the current study only employed 4 trials). Upon the first introduction to the task, participants were presented with an instructional recording saying: "Remember these letters, then write them out in the same order". Participants were then given a six-letter simultaneous spoken/written prompt that played a clear human-recording of the letter name while the letter itself appeared on the screen for 0.75 seconds per letter with a 0.25 seconds break between letter presentations. Afterwards, participants were given up to 24 seconds to recall all six letters by writing their responses on the screen with their finger or a stylus. Before writing, they were required to click on the input box for that letter. Previous research with similar tasks and similar populations has been conducted in a controlled laboratory setting (e.g., Elvevåg et al., 2001) with an experimenter closely observing and thus attempting to ensure that participants recalled the letters in the order specified.

2.3. Establishing the ground truth for letter responses

From the serial recall task, a series of screen-touch strokes (coordinate points and the corresponding timestamps) were collected to obtain a data series as well as the resulting PNG image files of the handwritten letters. Participants ranged from producing none to all six letters (see Supplementary figure).

Although the prompt letters were presented in uppercase, some participants responded with lowercase letters but these were rare (1.1% of all the various letter responses). For current purposes, lowercase letters are comparable to uppercase and so converted to uppercase. Three people transcribed the handwritten PNG files into typed letter sequences. Every letter was transcribed twice by two different people with roughly 98% agreement. A final judgment was made among the discrepancies but these inconsistencies occurred very rarely. Overall, the objective (i.e., provable) data for the various PNG responses (namely the letters the participants intended to write, which we term 'ground truth') could be identified reliably. (Although not the focus of this study, in section 4.1 below we report that when using machine learning to recognize these handwritten letters and transcribe them we were able to achieve accuracies of at least 95% which will in the future enable full automation of this process in technological implementations of STM assessment).

2.4. Establishing the level of compliance in participants' responses

Although remote testing did not require compliance with task instructions, because the positions of touches were recorded with milliseconds timestamps (constrained by the 60Hz sampling rate of the touchscreens) in a structured XML format (see Supplementary figure), the data revealed the order of recall through touch recognition. 'Compliant participants' were operationally defined as those who, after completion of the four serial recall tasks in a session, had no more than one entirely empty response sequence, namely six omission errors. There were 25 participants (22 patients, 3 nonpatients) who thus were excluded from further analysis. Information regarding demographics and primary diagnoses of the resulting samples are in Table 1. In addition, 17 empty responses (14 from patients, 3 from nonpatients) and 33 responses (26 from patients, 7 from nonpatients) that took longer than the allotted 24 seconds were removed. The remaining data consisted of 1,098 responses (540 from patients, 558 from nonpatients). All the responses were manually checked for irregularities to ensure that the included participants had exerted a reasonable effort.

3. Results

3.1. Order compliance analysis

To establish if participants complied with the instructions of recalling in the original order in which the items were presented (1, 2, 3, 4, 5, 6), the rate of correct versus incorrect response orders was calculated. It is noteworthy however that the correct order of recall does not

Table 1
Characteristics of compliant patients with substance use diagnoses and non-patient volunteers.

Characteristics		Patients (n = 71)	Nonpatients (n = 103)
Age (Years) (SD)		35.2 (9.9)*	19.9 (1.9)*
Gender	Female	0	23
	Male	71	80
Ethnicity	Caucasian/White	35	77
	African American/ Black	31	10
	Other	2	13
	NA	3	3
Education (Years)	<12	23	-
	12	31	-
	(12-13]	12	103
	16	2	-
	NA	3	-
Primary Diagnosis	None	-	103
	Substance Use	28	-
	Depression	20	-
	Anxiety	10	-
	Mood Disorder	5	-
	Bipolar	4	-
	PTSD	2	-
	Schizophrenia	2	-

* Welch's t-test $p < 0.001$

necessarily imply that the response itself was the correctly recalled letter, only that the input boxes were activated in the expected order. Not surprisingly, nonpatient participants recalled the six letter sequence in the correct order much more frequently (87.4%) than patients (51.3%). The majority (96.9%) of the nonpatient group recalled all six characters, as opposed to 56.1% of the patients.

3.2. Accuracy analysis based upon compliant responses only

Serial position effects are probably the most robust findings in STM research, characterized by the familiar U-shaped performance profile that reflects recency and primacy effects. Therefore, to establish the proof-of-concept in the current task design, the accuracy of recall was examined as a function of serial position. First, this was conducted as a function of all the data from compliant performances, then second from only data obtained in the first session of each participant. The accuracy of each group was computed by dividing the total number of correct letters by the total number of letters. A correct letter was defined as a prompt letter that was recalled in the correct position.

3.2.1. Overall accuracy

The average accuracy per serial position was graphed with each group as different trend lines. To observe the effect of phonological similarity, performance from the confusable and non-confusable letters were graphed both separately and combined.

As expected, primacy and recency effects were evident for both groups (Fig. 1) (e.g., Deese, 1957; Jahnke, 1965). Additionally, there was a clear phonological similarity effect (Conrad & Hull, 1964) with visibly poorer performance for confusable letters than for non-confusable letters. Patients' performances were consistently poorer. Importantly, these effects (primacy, recency and phonological similarity effect) were present for both groups (and - although we did not include an original version - were comparable to Elvevåg et al., 2001), and thus serve as proof-of-concept for such a technological implementation of a traditional task.

3.2.2. Accuracy analysis of the first session

To avoid the possibility of there being an advantage for those who had participated in multiple sessions, the data shown in Fig. 1 are from the first session only for all compliant responses. Within the nonpatient group there was only one person who participated in all three sessions, while in the patient group there were 22 people who completed all three sessions. This was a significant difference that could potentially influence the results. Therefore, choosing only the first sessions of all participants ensures that there is no advantage attributed because of a difference in practice.

When comparing accuracy from the first session only (Fig. 2) to overall accuracy across all data (Fig. 1), the serial-position effect is evident throughout, and patients' performance is poorer overall. Furthermore, Fig. 2 shows that patients' performance is slightly poorer than in Fig. 1, while the nonpatient group's data is similar across both analyses.

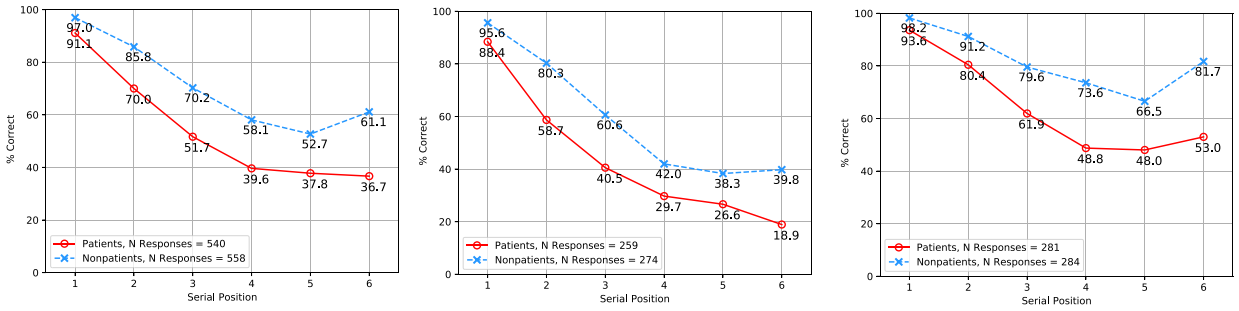
Accuracy, collapsed across phonologically confusable and non-confusable letters, is consistent with findings from traditional tests in patients and control participants (Elvevåg et al., 2001; see Table 2).

To establish if the phonological similarity effect was robust and similar across both participant groups, as seen in Fig. 2, a repeated measures analysis of variance (ANOVA) with group (patients, non-patients) as the between-subject factor and phonological similarity (confusable, non-confusable) as the within-subject factor was conducted. As expected, the phonological similarity effect was robust ($F(1, 170) = 122.5$, $p < 0.001$), with the confusable letters being recalled significantly less accurately than the non-confusable letters. Also here patients performance was overall less accurate ($F(1, 170) = 58.5$, $p < 0.001$) (Table 2) but the recall pattern was similar across both participant groups (Group × Phonological Similarity; $F(1, 170) = 1.0$, $p = 0.31$). To examine the effect of item position on accuracy, a repeated measures ANOVA with group (patients, nonpatients) as the between-subject factor and phonological similarity (confusable, non-confusable) and serial position (1, 2, 3, 4, 5, 6) as within-subject factors was conducted. There was a main effect of group and phonological similarity (as mentioned above), as well as a main effect of position ($F(5, 850) = 200.4$, $p < 0.001$) likely because of the classic U effect. Interestingly, there was a Group × Position interaction ($F(5, 850) = 8.1$, $p < 0.001$), which might relate to the less obvious recency effect for patients, as well as their difficulty in recalling the last few letters. Additionally, there was a Phonological Similarity × Position interaction ($F(5, 850) = 15.0$, $p < 0.001$), as well as a Group × Phonological Similarity × Position interaction ($F(5, 850) = 3.6$, $p = 0.003$).

Compared with findings from a previously published traditional pen and paper version in a controlled laboratory setting (Elvevåg et al., 2001), the current data seem remarkably similar and provide support for the clinical viability of a technological implementation of this traditional task. The graphs and ANOVA analyses show the classic serial-position effect for recall accuracy, as well as the phonological similarity effect of recall of confusable letters being less accurate than non-confusable letters for both groups. Analyses of the data from the first sessions demonstrate that despite using significantly fewer trials than the more traditional task (4 vs. 20) and implementing the task differently, a good quality dataset was nonetheless produced as evidenced by the signature serial position and phonological similarity effects.

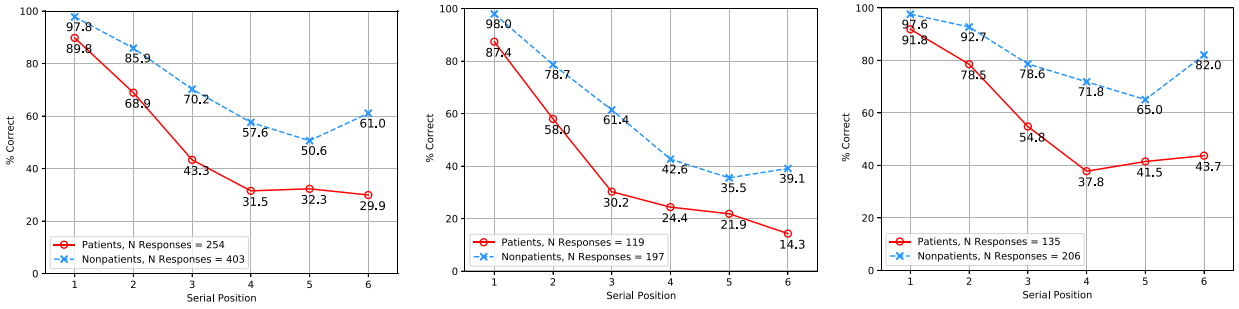
3.3. Timing analysis

For over a century it has been speculated that time may be the critical factor that affects short-term memory performance, and indeed time has been integrated into the design of numerous controlled experiments (e.g., Cowan & AuBuchon, 2008; Jonides et al., 2008; Oberauer & Lewandowsky, 2008; Thomas et al., 2003). However, to our knowledge no study has explicitly examined the underlying temporal dynamics and the manner in which recall times affect recall accuracy, yet this is now possible with high precision using technology to collect and examine written (or spoken) responses. Also, since it stands to reason that clinical populations will display behavioral performance



(a) collapsed (b) confusable (c) non-confusable

Fig. 1. Overall accuracy as a function of the serial position of the six letters (Panel a) collapsed across phonological similarity, and then separately for (Panel b) confusable letters and (Panel c) non-confusable letters. (a) collapsed (b) confusable (c) non-confusable.



(a) collapsed (b) confusable (c) non-confusable

Fig. 2. Accuracy by serial position for six letters: data from the first session only (Panel a) collapsed across phonological similarity, and then separately for (Panel b) confusable letters and (Panel c) non-confusable letters. (a) collapsed (b) confusable (c) non-confusable.

Table 2
Overall performance (% accuracy) of both groups in the first sessions for both phonologically confusable and non-confusable sets of letters collapsed and then separately.

Group	Collapsed	Confusable	Non-Confusable
Patients	49	39	58
Nonpatients	71	59	81

characterised by slower reaction times, a framework with which to explore dynamics promises to be useful. Therefore, this study sought to examine recall times explicitly, and explore the relationships they may have with recall accuracy, error, and serial position. The well-known primacy, recency, and phonological-similarity effects were thus expanded beyond accuracy to also include performance in time.

The components of time used by compliant participants (71 patients, 103 nonpatients) during serial recall were deconstructed and illustrated in Fig. 3. The novel mechanism developed to analyse the data for the current study includes four components: first, was the design of 'the letter input box'; second, the participant needed to 'click the letter input box' to trigger the timing and indicate a mental switch was made to begin the recall process of the letter; third was the timing of the first touch or tap; and fourth was the timing of the last touch or tap (see Fig. 3).

Thus, if a participant left an input box blank (an output omission error - see Henson, 1998; Elvevåg et al., 2001), the retrieval, action, and elapsed times were all registered with a null value, and as there were no data for these cases, they were excluded from subsequent analyses. As patients more frequently left input boxes blank, the resulting dataset for timing analysis was smaller for patients than for nonpatient participants.

The average retrieval, action, and elapsed times are graphed as a

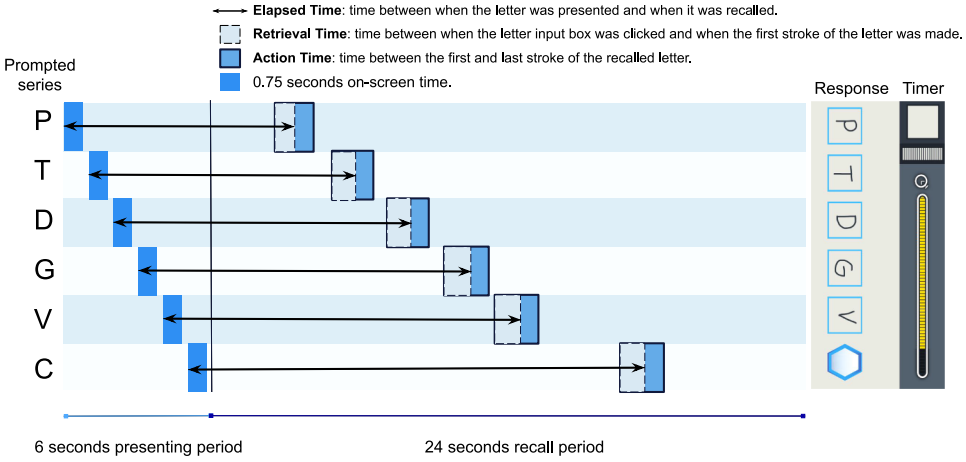
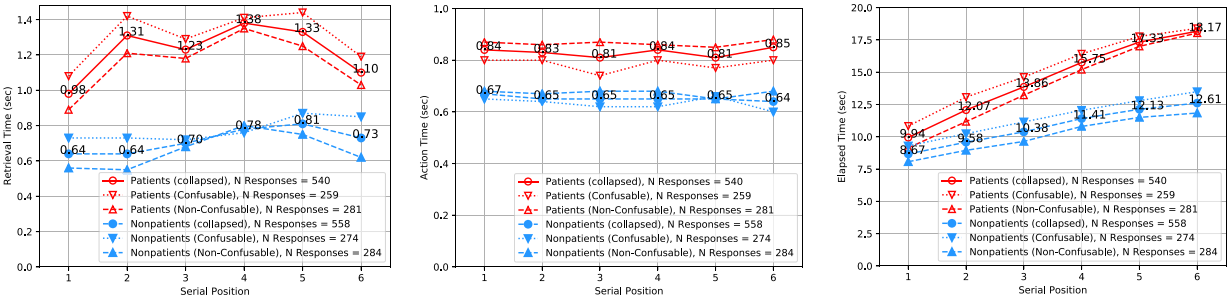


Fig. 3. Diagram of three novel time variables deconstructed from total response time. The 'Elapsed Time' is the time from letter presentation until recall. The 'Retrieval Time' is computed as the time between the click-activation of the letter box and the beginning of the first stroke of writing the letter. The 'Action Time' is computed as the time between the beginning of the first stroke of the recalled letter and the end of the last stroke of writing this letter.



(a) retrieval time (b) action time (c) elapsed time

Fig. 4. Time spent per position for retrieval (Panel a), action (Panel b), and elapsed (Panel c) time. In each panel each position, the middle point is the time value collapsed across phonological similarity, and then other two points are separately for confusable letters and non-confusable letters. (a) retrieval time (b) action time (c) elapsed time

function of serial position in Fig. 4. Three repeated measures ANOVA tests were performed, with retrieval, action, and elapsed times as the three different dependent variables, and with group (patients, non-patients) as the between-subject factor, and position (1, 2, 3, 4, 5, 6) as the within-subject factor. For 'retrieval time', there was a significant main effect of group ($F(1, 162) = 216.16$, $p < 0.001$), and position ($F(5, 810) = 8.50$, $p < 0.001$). There was also a significant Group × Position interaction ($F(5, 810) = 3.65$, $p = 0.003$). For 'action time', there was a main effect of group ($F(1, 162) = 54.94$, $p < 0.001$), but no main effect of position ($F(5, 810) = 0.59$, $p = 0.71$), nor a significant interaction between Group × Position ($F(5, 810) = 0.88$, $p = 0.50$). For 'elapsed time', there was a main effect of Group ($F(1, 162) = 251.64$, $p < 0.001$), as well as Position ($F(5, 810) = 574.69$, $p < 0.001$). Additionally, there was a significant interaction between Group × Position ($F(5, 810) = 75.85$, $p < 0.001$), which is consistent with patients spending more time in action and retrieval time. When phonological similarity (confusable, non-confusable) was added as a within-subject factor, there was a main effect for retrieval, action, and elapsed time ($p < 0.001$), but no interaction with other factors. When comparing the curves plotted from the confusable and non-confusable letter data in Fig. 4(a), it is clear that not surprisingly phonologically confusable letters resulted in longer retrieval times than non-confusable letters. However, Fig. 4(b) shows that non-confusable letters resulted in longer action times than confusable letters, which may be due to the higher number of strokes required to write non-confusable letters (e.g., 1.7 strokes on average for confusable letters versus 2.2 strokes on average for non-confusable letters). Fig. 4(c) shows that a sequence of phonologically confusable letters resulted in longer elapsed times than a sequence of non-confusable letters.

In summary, patients were consistently slower than nonpatients for all three different times. For 'retrieval time', there was a slight inverted U-shaped curve, showing the presence of the primacy and recency effects in a time domain for both groups (Fig. 4(a)). This was especially prominent for patients, causing a significant repeated measures ANOVA interaction between group and serial position. For 'action time', both groups displayed a consistent speed for all six serial positions, suggesting that letter-writing may be relatively constant for participants. For 'elapsed time', both groups displayed longer lapses of time as the serial position increased. The elapsed time differential showed a greater absolute increase by position for patients, who were generally slower, than for nonpatient participants. These novel analyses (Figs. 4) show the significant value of the novel measures of memory 'retrieval time'.

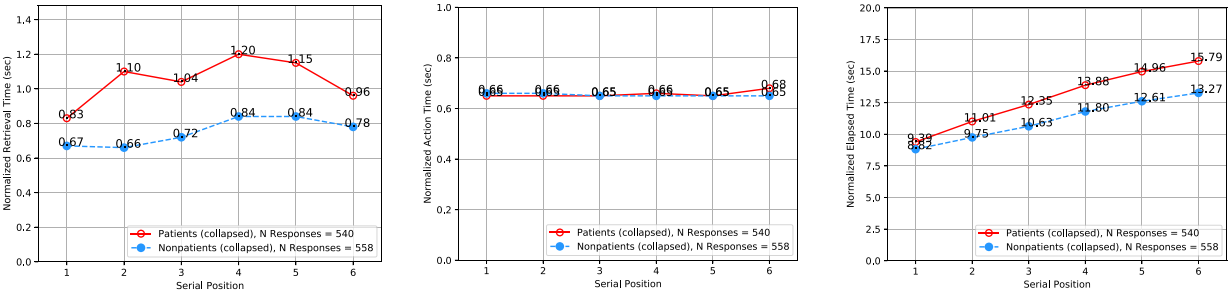
3.3.1. Normalization by action speed

The overall average action time for patients was 0.830 seconds and 0.653 for nonpatient participants. In order to comment most cautiously regarding patients' slowness, normalization is necessary to remove the general effect of slower action times and to gain an understanding of other cognitive mechanisms that may account for group differences. Retrieval (R), action (A) and elapsed (E) times were normalized using the average action speed ($\bar{A}$) across all positions per response for all participants:

$$\text{e.g.,}\ \hat{R} = R \cdot 0.653/\bar{A},$$
$$\hat{A} = A \cdot 0.653/\bar{A}, \text{and}$$
$$\hat{E} = (E - (7 - p)) \cdot 0.653/\bar{A} + (7 - p),\ p = 1, \ldots, 6.$$

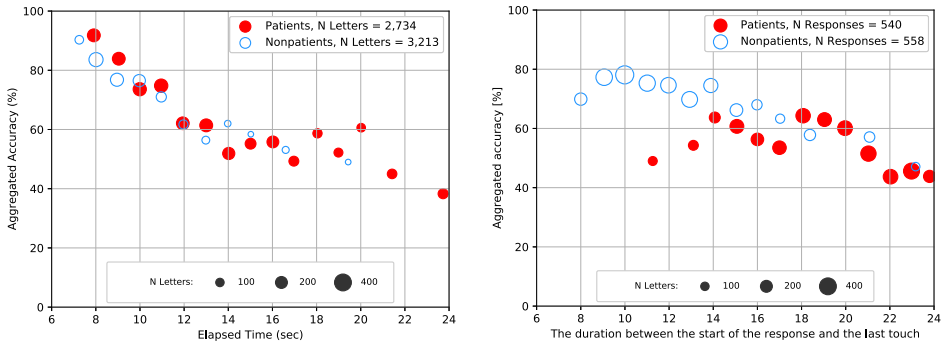
Since presentation time is a constant and not affected by slowness, the elapsed time within the first six seconds was subtracted from elapsed time before being added back after multiplying by the ratio. If the group differences in serial positions were caused by patients' overall slower action speeds, one would expect to see the same group values in normalized data (such as Fig. 5(b)), but if not then there must be other factors driving the differences.

It is important to note that the differences between groups for both retrieval and elapsed times after normalization were smaller (Fig. 5(a), 5(c)), thus indicating a central role of the naturally slower action speed of patients. However, the normalized results were not similar, which suggests that further unaccounted factors may underlie group differences.



(a) normalized retrieval time (b) normalized action time (c) normalized elapsed time

Fig. 5. Retrieval, action, and elapsed time versus serial position after normalized by action time to remove the general effect of action speed. (a) normalized retrieval time (b) normalized action time (c) normalized elapsed time



(a) elapsed time vs. accuracy (b) response duration vs. accuracy

Fig. 6. The relationships between the presentation elapsed time and recall accuracy (Panel a), and between the total response duration per recall and accuracy (Panel b), with dot size representing number of letters. (a) elapsed time vs. accuracy (b) response duration vs. accuracy.

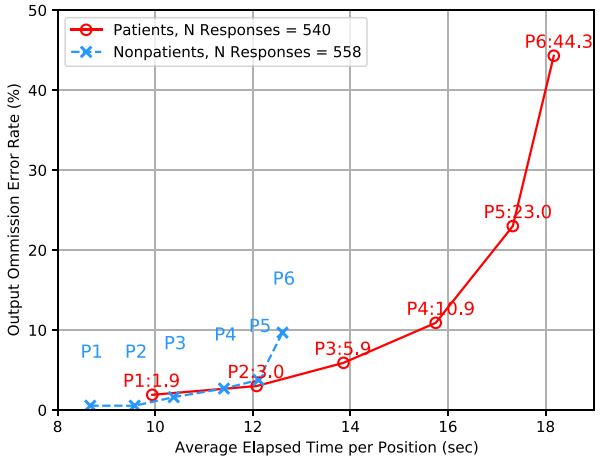


Fig. 7. Elapsed time by output omission error rate by output position. P1, P2, …, P6 stand for the serial position.

3.3.2. *Temporal dynamics on recall performance*

Average accuracy was graphed by a condensed elapsed time range to show the relationship between elapsed time and accuracy (Fig. 6(a)). The elapsed times of the letters were sorted from the shortest to longest before being averaged. Each point in Fig. 6(a) represents the average accuracy (the y-axis) of at least 100 letters by the rounded integer second of elapsed time for those letters. If there were fewer than 100 letters for an integer second, the data would continually be averaged with the following rounded integer seconds until there were more than 100 letter values. The elapsed time value for this group of letters would be the average of their elapsed time (the x-axis). In total, there were 2,734 letters for patients and 3,213 letters for nonpatient participants. From Fig. 6(a), it is apparent that there was a strong relationship between time and accuracy: as elapsed time increased, the accuracy of recall decreased. There was no obvious difference between the correlation of patients: 0.87, and nonpatient participants: 0.94, with a correlation of 0.90 for all the data taken together. The principal difference seen in the data is that the nonpatient group generally had shorter elapsed times while patients' data displayed a more distributed elapsed time range extending to include longer times. There was a clear relationship between patients' poorer performance and their tendency to recall slower. Although Lewandowsky et al. (2004) argued that time *per se* does not cause forgetting in STM serial recall, the current results indicate that the elapsed times may be a good indicator of the performance accuracies. The similarity of results between the groups shows that patients with varying levels of illness are nonetheless able to perform at a similar accuracy level to nonpatients for the same amount of elapsed time. However, patients made 12% more output omission errors which were not used in this analysis as there was no elapsed time when these errors occurred, hence fewer letters for patients (Fig. 6(a)).

Fig. 6(b) displays an aggregate time variable: Response duration, which is the total response time a participant spent recalling six letters. Taking $\boldsymbol{n}$ as the total number of letters recalled in a single response, response duration = sum(n action times, n retrieval times, n pauses between actions), or the time from the first box click to end of the last letter stroke. Response duration was graphed as a function of the aggregated accuracy in Fig. 6(b). Unlike Fig. 6(a), output omission errors were included in each response's accuracy although they may have occurred after the last stroke. The data were analysed in a similar manner to that in Fig. 6(a), with at least 100 letters per data point. Fig. 6(b) shows that on average, patients tended to spend more time recalling than nonpatient participants. In the nonpatient groups, the more time spent, the lower recall accuracy. However, although there was no clear correlation for patients, their generally slower recall was related to their poorer performance. When the duration was less than 13.5s, patients made 14% more output omission errors and 10% more movement errors than the nonpatient group, which resulted in lower accuracy at those elapsed times. Most patients with shorter response durations did not recall the last letters, resulting in lower accuracy due to those output omission errors.

As displayed in Fig. 7, the relationship between the average elapsed time (x-value) and the output omission error rates (y-value) for each serial position per group was examined further. Since omission errors do not have any elapsed time values because no stroke was made, an assumption was made that if they were to occur, the average elapsed times of recalled letters from the same group and position could be used as a proxy for the elapsed times for these omission errors.

Fig. 7 illustrates that with longer elapsed time there were more output omission errors. Additionally, for both groups, there was a significant increase in the percentage of output omissions between the longest elapsed times: For patients there was a 12.0% absolute increase between 15.7s (position 4) and 17.3s (position 5), and 21.3% absolute increase between 17.3s (position 5) and 18.2s (position 6), and in the nonpatient group there was a 5.9% absolute increase between an elapsed time of 12.1s (position 5) and 12.6s (position 6). The significant increase (in both groups) between positions 5 and 6 may be attributed to Miller's well-known 7 ± 2 observation (Miller, 1956). As patients were generally slower and displayed longer elapsed times, the higher percentage of output omission errors seem clearly related to the longer time spans between letters.

4. Discussion

This self-administered smart-device version of a STM task was practical, reliable and generated data comparable to data produced from traditionally administered tasks in studies conducted in controlled laboratory settings. Patients with mental illness were generally able to

perform the task and displayed robust STM effects in a comparable manner to nonpatient participants, but with the general reductions in speed and accuracy expected in this patient population. The qualitatively similar performance pattern in patients suggests that the underlying mechanisms for recall are similar despite the illness, medication or overall slower response pace.

The implementation of new technology and specific designs implemented in the app make it possible to collect, extract and analyze participants' exact retrieval, action, and elapsed times. It is clear (and not surprising) that patients inherently took a longer time for the entire recall process than nonpatients did. Comparing the elapsed time per serial position to the omission errors per serial position showed that the longer elapsed time between recall of letter was correlated with the increased likelihood of omitting a letter during recall. The primacy and recency effects are traditionally reflected in performance accuracy, however in this study the effects were also observed in the time domain. The primacy and recency effects were evident in retrieval time as an inverted U-shaped curve, and a reliable phonological similarity effect was observed in retrieval time.

As operationalized in this study, memory 'retrieval time' is a novel metric that may be of value in future studies that seek to move beyond this proof-of-concept design to a more clinical study design. We found that the items presented earlier and later within a sequence took less time to retrieve, and the middle items took longer time to retrieve which is consistent with the concept that these middle items are more difficult to recollect. In addition, we found that phonologically similar letters took longer time to retrieve. These findings are all consistent with the traditional and well-established primacy, recency and phonological similarity effects in accuracy, but are here extended to variables expressed in the precise recall time.

4.1. Limitations

This study was a proof-of-concept study and not designed to probe differences between groups, and as such there are some noteworthy limitations. This study was not designed to make specific inferences about the role of mental illness on STM, and to do that future studies will have to control for key variables such as age and levels of education. Further, although the student volunteers were assumed to be healthy as compared to patients, they did not go through any mental illness screening. Indeed, the significant age difference between the two groups (see Table 1) may have impacted the results, as younger participants may of course be faster. Additionally, this study was conducted with patients whose behavioral responses are classically defined by variability. This translates to data which may not be as perfectly normally distributed as nonpatients participants (who standardly are studied in memory research). However, since the resulting patterns from the current analyses are remarkably similar to the existing literature, there is no *a priori* reason to be especially concerned. Furthermore, it has been previously found in similar paradigms that the complex transformation of data to meet the assumptions of parametric tests (e.g., normal distribution) nonetheless mirror the original non-normally distributed data (Elvevåg et al., 2001).

4.2. Full automation

The identity of the letter responses reported in this paper were ascertained by human judgments (i.e., by a subset of the authors). Recognizing letters from PNG image files is a classic machine learning problem, namely isolated handwritten letter recognition (e.g., Ciresan et al., 2011, 2012; Cohen et al., 2017). The deep learning techniques (such as convolutional neural networks, CNNs) can now achieve accurate recognition rates of 98% or better for handwritten letters (Ciresan et al., 2011). In our judgment, the human consensus letter recognition reported here can be taken as 'ground truth'. Our preliminary testing showed that we could achieve accuracies of 95% or more when we applied the same CNN-based deep learning architecture to these data. By applying this 95% level accuracy algorithm, all performance metrics in this report could be computed on current smart-devices and presented in real-time. Put simply, full automation of the analysis of the data from this task is viable (that is, from PNG image file to letter transcription to scoring).

4.3. Summary and conclusion

Data gathered through smart-devices from self-administered short-term memory tasks for sequential order were reliable. Patients' generally poorer performance was probably attributable to their slower cognitive processes, thus the omission of letters at later serial positions. A new mechanism was proposed to capture the actual memory retrieval time, and three new time variables were defined to provide a novel way to perform analyses. This study shows that primacy and recency effects are also present in a time domain, as represented by the inverted U-shaped curve in memory retrieval time. Also, the phonological similarity effect is evident in retrieval time. There is a strong correlation between longer elapsed time and lower accuracy for both groups, showing that cognitive speed of short-term memory processes affects accuracy, errors, and timing. Confirmation of the precision and reliability of data collected through a smart-device opens up opportunities for future studies that seek to distribute tasks to large populations and collect data that records multiple aspects of participants' recall. By applying recent technological advances, every interaction is recorded precisely thus enabling an extremely detailed investigation of performance and errors on a significantly larger scale than has been previously possible.

However, current memory tasks aim for difficulty levels to surpass the average person's ability, rather than the individual's. As technology advances, the implementation of adaptive learning for memory tasks – where the difficulty of the task assigned changes in response to past participant performances – should allow for a more sensitive and in-depth understanding of memory ability as each participant would be tested based on their personal threshold. Future studies could implement a design to control elapsed time to make elapsed time spent equivalent across groups so that performance differences can be studied in such conditions. While specific clinical disorders and the time variables were not deeply explored in this report, carefully designed experiments in the near future can explore the relationships between time variables and clinical conditions that affect short-term memory. Given the critical relationship of temporal dynamics with serial recall, this technological approach to data collection and analysis promises to be scientifically interesting in a variety of cortical disorders by enabling the parsing of overall slowing in reaction time in short-term memory performance. In psychiatry specifically, this detailed attention to the temporal dynamics in short term recall may be a helpful additional tool when monitoring the effects of medication.

Authors disclosures

Jian Cheng, Jared Bernstein, and Elizabeth Rosenfeld have a financial interest in Analytic Measures Inc., which owns the intellectual property in the software system described in the paper. The other authors have nothing to disclose.

Authors' contribution

All authors contributed substantially and meaningfully to this study and the final manuscript. BE, JCB, PWF and ASC for funding. BE, JCB, JC, PWF, ASC, and EPR planned the study and were involved in the smart-device app design. JC coded the app and the whole data collection platform. ASC led the participants' recruitment and the data collection phase. SC conducted data analysis under BE's supervision. Some analyses ideas were proposed by SC during the data analysis process. SC

wrote the first draft of the manuscript and used it to participate in the Regeneron Science Talent Search Competition in 2019. JC, BE, and TBH reviewed and edited the original draft to be suitable for publication. All authors have approved the final manuscript.

Ethics

The data were collected through a study approved by the Louisiana State University (LSU) Institutional Review Board (#3618). All participants provided informed consent and signed consent forms in accordance with the Helsinki declaration prior to participating.

Declaration of Competing Interest

The authors declare no conflicts of interest associated with this research study.

Acknowledgement

This research was funded by a grant from the Research Council of Norway to Brita Elvevåg (#231395). The funding source had no involvement in the study design, in the collection, analysis or interpretation of data, in the writing of the manuscript or in the decision to submit the article for publication.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2020.113494.

References

Analytic Measures Inc. (2018). dMSE. Retrieved from https://apps.apple.com/us/app/dmse/id1134008913 (Computer Software. Apple App Store, Version 1.1.5.).

Botvinick, M.M., Plaut, D.C., 2006. Short-term memory for serial order: a recurrent neural network model. Psychological Review 113 (2). https://doi.org/10.1037/0033-295X.113.2.201. 201–33.

Brown, G.D.A., Neath, I., Chater, N., 2007. A temporal ratio model of memory. Psychological Review 114, 539–576. https://doi.org/10.1037/0033-295X.114.3.539.

Chandler, C., Foltz, P.W., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Cohen, A.S., Holmlund, T.B., Elvevåg, B., 2019. Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In: Niederhoffer, K., Hollingshead, K., Resnik, P., Resnik, R., Loveys, K. (Eds.), Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. Minneapolis, Minnesota, USA. 2019. pp. 137–147. https://aclweb.org/anthology/volumes/proceedings-of-the-sixth-workshop-on-computational-linguistics-and-clinical-psychology/.

Chandler, C., Foltz, P.W., Cheng, J., Cohen, A.S., Holmlund, T.B., Elvevåg, B., 2020. Predicting Self-Reported Affect from Speech Acoustics and Language. In: Proceedings of the LREC 2020 Workshop on: Resources and Processing of Linguistic, Para-linguistic and Extra-linguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments (RaPID-3), pp. 9–14. https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/RaPID3book.pdf.

Cheng, J., Bernstein, J., Rosenfeld, E., Foltz, P.W., Cohen, A.S., Holmlund, T.B., Elvevåg, B., 2018. Modeling self-reported and observed affect from speech. In: Proc. interspeech 2018, pp. 3653–3657. https://doi.org/10.21437/Interspeech.2018-2222.

Ciresan, D.C., Meier, U., Schmidhuber, J., 2012. Multi-column deep neural networks for image classification. In: Proceedings of the 25th IEEE conference on computer vision and pattern recognition (CVPR 2012), pp. 3642–3649. https://doi.org/10.1109/CVPR.2012.6248110.

Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J., 2011. Convolutional neural network committees for handwritten character classification. In: ICDAR (pp. 1135–1139). IEEE Computer Society, https://doi.org/10.1109/ICDAR.2011.229.

Cohen, A.S., Fedechko, T.L., Schwartz, E.K., Le, T.P., Foltz, P.W., Bernstein, J., Cheng, J., Holmlund, T.B., Elvevåg, B., 2019. Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. Journal of Abnormal Psychology 128 (2), 97–105. https://doi.org/10.1037/abn0000397.

Cohen, G., Afshar, S., Tapson, J., van Schaik, A., 2017. EMNIST: an extension of MNIST to handwritten letters. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 2921–2926. https://doi.org/10.1109/IJCNN.2017.7966217.

Conrad, R., Hull, A.J., 1964. Information, acoustic confusion and memory span. British Journal of Psychology 55 (4), 429–432. https://doi.org/10.1111/j.2044-8295.1964.tb00928.x.

Cowan, N., AuBuchon, A.M., 2008. Short-term memory loss over time without retroactive stimulus interference. Psychonomic Bulletin & Review 15 (1), 230–235. https://doi.org/10.3758/PBR.15.1.230.

Deese, J., 1957. Serial organization in the recall of disconnected items. Psychological Reports 3, 577—582. https://doi.org/10.2466/pr0.1957.3.3.577.

Ebbinghaus, H., 1885. Über das gedächtnis [on memory]. Duncker & Humblot, Leipzig, Germany.

Elvevåg, B., Weinberger, D.R., Goldberg, T.E., 2001. Short-term memory for serial order in schizophrenia: a detailed examination of error types. Neuropsychology 15, 128–135. https://doi.org/10.1037/0894-4105.15.1.128.

Farrell, S., 2006. Mixed-list phonological similarity effects in delayed serial recall. Journal of Memory and Language 55, 587–600. https://doi.org/10.1016/j.jml.2006.06.002.

Farrell, S., Lewandowsky, S., 2002. An endogenous distributed model of ordering in serial recall. Psychonomic Bulletin & Review 9 (1), 59–79. https://doi.org/10.3758/BF03196257.

Farrell, S., Lewandowsky, S., 2004. Modelling transposition latencies: constraints for theories of serial order memory. Journal of Memory and Language 51 (1), 115–135. https://doi.org/10.1016/j.jml.2004.03.007.

Henson, R., 1998. Short-term memory for serial order: the start-end model. Cognitive Psychology 36, 73–137. https://doi.org/10.1006/cogp.1998.0685.

Holmlund, T.B., Foltz, P.W., Cohen, A.S., Johansen, H., Sigurdsen, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E., Elvevåg, B., 2019a. Moving psychological assessment out of the controlled laboratory setting: practical challenges. Psychological Assessment 31 (3), 292–303. https://doi.org/10.1037/pas0000647.

Holmlund, T.B., Cheng, J., Foltz, P.W., Cohen, A.S., Elvevåg, B., 2019b. Updating verbal fluency analysis for the 21st century: Applications for psychiatry. Psychiatry Research 273, 767–769. https://doi.org/10.1016/j.psychres.2019.02.014.

Holmlund, T.B., Chandler, C., Foltz, P.W., Cohen, A.S., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Elvevåg, B., 2020. Applying speech technologies to assess verbal memory in patients with serious mental illness. npj Digit. Med 3, 33. https://doi.org/10.1038/s41746-020-0241-7.

Jahnke, J.C., 1965. Primacy and recency effects in serial-position curves of immediate recall. Journal of Experimental Psychology 70, 130–132. https://doi.org/10.1037/h0022013.

Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., Moore, K.S., 2008. The mind and brain of short-term memory. Annual Review of Psychology 59 (1), 193–224. https://doi.org/10.1146/annurev.psych.59.103006.093615.

Kahana, M.J., Jacobs, J., 2000. Interresponse times in serial recall: effects of intraserial repetition. Journal of Experimental Psychology. Learning, Memory, and Cognition 26 (5), 1188–1197. https://doi.org/10.1037/0278-7393.26.5.1188.

Lewandowsky, S., Duncan, M., Brown, G.D.A., 2004. Time does not cause forgetting in short-term serial recall. Psychonomic Bulletin & Review 11 (5), 771–790. https://doi.org/10.3758/BF03196705.

Maybery, M.T., Parmentier, F.B., Jones, D.M., 2002. Grouping of list items reflected in the timing of recall: implications for models of serial verbal memory. Journal of Memory and Language 47, 360–385. https://doi.org/10.1016/S0749-596X(02)00014-1.

Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. The Psychological Review 63 (2), 81–97. https://doi.org/10.1037/h0043158.

Nipher, F.E., 1876. On the distribution of numbers written from memory. Transactions of the Academy of St. Louis 3, 79–80.

Oberauer, K., 2003. Understanding serial position curves in short-term recognition and recall. Journal of Memory and Language 49, 469—483. https://doi.org/10.1016/S0749-596X(03)00080-9.

Oberauer, K., Lewandowsky, S., 2008. Forgetting in immediate serial recall: decay, temporal distinctiveness, or interference? Psychological Review 115 (3), 544–576. https://doi.org/10.1037/0033-295X.115.3.544.

Page, M., Norris, D., 1998. The primacy model: a new model of immediate serial recall. Psychological Review 105 (4), 761–781. https://doi.org/10.1037/0033-295X.105.4.761-781.

Thomas, J.G., Milner, H.R., Haberlandt, K.F., 2003. Forward and backward recall: different response time patterns, same retrieval order. Psychological Science 14 (2), 169–174. https://doi.org/10.1111/1467-9280.01437.