Psychonomic Bulletin & Review (2023) 30:308–316
https://doi.org/10.3758/s13423-022-02171-4

BRIEF REPORT

Machine-learning as a validated tool to characterize individual differences in free recall of naturalistic events

Xinxu Shen[1] · Troy Houser[2] · David V. Smith[1] · Vishnu P. Murty[1]

Accepted: 24 August 2022 / Published online: 9 September 2022

Abstract

The use of naturalistic stimuli, such as narrative movies, is gaining popularity in many fields, characterizing memory, affect, and decision-making. Narrative recall paradigms are often used to capture the complexity and richness of memory for naturalistic events. However, scoring narrative recalls is time-consuming and prone to human biases. Here, we show the validity and reliability of using a natural language processing tool, the Universal Sentence Encoder (USE), to automatically score narrative recalls. We compared the reliability in scoring made between two independent raters (i.e., hand scored) and between our automated algorithm and individual raters (i.e., automated) on trial-unique video clips of magic tricks. Study 1 showed that our automated segmentation approaches yielded high reliability and reflected measures yielded by hand scoring. Study 1 further showed that the results using USE outperformed another popular natural language processing tool, GloVe. In Study 2, we tested whether our automated approach remained valid when testing individuals varying on clinically relevant dimensions that influence episodic memory, age, and anxiety. We found that our automated approach was equally reliable across both age groups and anxiety groups, which shows the efficacy of our approach to assess narrative recall in large-scale individual difference analysis. In sum, these findings suggested that machine learning approach implementing USE is a promising tool for scoring large-scale narrative recalls and perform individual difference analysis for research using naturalistic stimuli.

Keywords Naturalistic stimuli · Machine learning · Episodic memory and recall

The use of naturalistic stimuli, such as movie clips, has yielded significant advances in fields characterizing memory, affect, and decision-making. Naturalistic stimuli provide the benefit of mimicking real-world situations and also provide the opportunity to elicit strong emotional states (Saarimäki, 2021; Sonkusare et al., 2019). Further, naturalistic stimuli improve our understanding of episodic memory by allowing participants to freely recall these complex stimuli during test (J. Chen et al., 2017; J. Chen et al., 2016; Coutanche et al., 2020; Ren et al., 2018; St-Laurent et al., 2016). However, the analysis of narrative recall of naturalistic events requires time-consuming, hand-scoring approaches which are prone to systematic bias. These limitations make it difficult to study narrative recall of naturalistic stimuli and preclude the ability to conduct large-scale individual difference analyses. Therefore, new methods are needed to automate the scoring and analysis of narrative recall. Here, we tested the efficacy of utilizing natural language processing (NLP) tools to score narrative recall.

Narrative recall paradigms are one of the most common measures of memory for naturalistic stimuli (J. Chen et al., 2017; J. Chen et al., 2016; Coutanche et al., 2020; Scheurich et al., 2021). During narrative recall, participants are asked to recall as much information as possible without any explicit associative cues beyond the title of the narrative. While narrative recall provides rich memory information, the procedure of scoring narrative recall data is often idiosyncratic. Scoring narrative recalls usually involves training researchers, partitioning participants' narrative recall responses and assessing accuracy for the recalled segments (J. Chen et al., 2016; Coutanche et al., 2020; Silva et al., 2019). Each step is time-consuming and could vary both within and across research groups. In our own research procedures, training researchers to reliably score narrative recalls can take weeks, and segmenting and

✉ Vishnu P. Murty
vishnu.murty@temple.edu

[1] Department of Psychology, Temple University, 1701 N 13th St, Philadelphia, PA 19122, USA

[2] Department of Psychology, University of Oregon, Eugene, OR 97403, USA

coding narrative recalls can take weeks to months, depending on the size of the dataset.

NLP tools are gaining popularity in text categorization and semantic similarity analysis (Cer et al., 2018; Devereux et al., 2013; Naspi et al., 2021; Pennington et al., 2014; Wang et al., 2018; Zhu et al., 2017), which makes them promising tools to automate the process of characterizing narrative recall. NLP tools, such as Global Vectors for Word Representation (GloVe), latent semantic analysis (LSA) and the Universal Sentence Encoder (USE), have been developed for systematically quantifying the relative meanings across words. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. GloVe has been shown to outperform LSA (Deerwester et al., 1990) and other semantic models on several validation tests (Pennington et al., 2014). While GloVe could be useful for scoring narrative recalls, the approach is trained at the word level, precluding the ability to code meaning at the sentence level which may be more relevant for narrative recall. An alternative approach utilizes the Universal Sentence Encoder (USE), which is a natural language processing tool that embeds not only words but phrases and sentences (Cer et al., 2018). In the context of narrative recall, we predict that USE might provide more accurate representations of memories conveyed through narrative recall.

Here, we designed two studies to test the reliability of NLP tools in scoring narrative recalls. In both studies, participants watched short magic clips (Ozono et al., 2020) on the first day, and after a 24-hour delay, participants completed a memory test to recall as much information as they could from the magic clips they watched on the previous day. In Study 1, we compared the reliability of two NLP tools, Global Vectors for Word Representation (GloVe) and the Universal Sentence Encoder (USE), against hand-scoring approaches to characterize free recall of short video clips. In Study 2, we extended these analyses to determine if our automated scoring procedures remain reliable when addressing clinically relevant populations (i.e., age and anxiety), to determine if they were resilient to individual differences in episodic memory. The goals of the current studies are to (1) test the reliability of using USE and GloVE to automatically score narrative recalls, (2) test the reliability of using these methods to automatically score large-scale individual difference data, and (3) develop an individual difference analysis toolbox that can be used for testing the effect of curiosity on memory for clinically relevant populations.

Study 1

Materials and methods

Participants

Twenty participants (ages 18–24) were recruited from Temple University via SONA as part of the subject pool for the Department of Psychology (https://www.sona-systems.com/default.aspx). Temple University's Institutional Review Board approved study materials and procedures. All participants provided informed consent and were compensated for their time with course credits.

Stimuli

Our task involved participants watching short magic clips. Twenty magic clip stimuli were drawn from the Magic Curiosity Arousing Tricks stimuli set (Ozono et al., 2020). To ensure sufficient interstimulus variability, which prevents overlapping in memories and ensures good quality of the recall data, we selected video stimuli with different phenomena categories, materials, lengths, and curiosity ratings. Magic clips that were scored above or below one standard deviation of average curiosity ratings were grouped as high- or low-curiosity stimuli, respectively. Average curiosity rating for high-curiosity video stimuli was 6.32 (out of 10), and average curiosity rating for low-curiosity stimuli was 4.86 (out of 10). Length of magic clips were matched for high-curiosity stimuli and low-curiosity stimuli. Average length of high-curiosity stimuli was 30 seconds and average length of low-curiosity stimuli was 28 seconds. Phenomena categories included transportation, color change, restoration, take one and other. No more than three videos were from the same phenomena categories, and we matched phenomena categories for high and low curiosity video clips. Magic clips included as high curiosity stimuli were S32, Trick7_Long, K12, S9, K4, H16, H11, K2, S16, and S10. Magic clips included as low curiosity stimuli were S15, S30, K24, H36, Trick14_Long, S21, H2, K25, S3_Short, and H20.

Procedure

On day one, participants watched the twenty magic clips with one of the three created random orders on Qualtrics. Participants first were instructed to click a button to start the first magic clip watching. The magic clip auto-played after participants clicked the button and participants were not able to pause or replay the magic clips during watching. After each magic clip ended, the page auto-advanced to the next page, where participants were instructed to answer two curiosity ratings ("How much you'd like to see a similar magic clip on YouTube" and "How surprised you are at the magic clip") on a 0–5 Likert scale about each magic clip. The rating task was self-paced and not relevant for this paper. After they finished the rating task, they were instructed to click a button to watch the next magic clip. After a 24-hour delay, participants completed a narrative recall task of magic clips they watched on the previous day. During the task, participants were presented with a screenshot of a magic clip (at 1s of the video) they watched on the first day and were instructed to write in one text box a step-by-step recreation of the magic clips based on their memory of the screenshot. We decided to characterize our scoring approach on written recalls to be more amenable to large scale data collection samples, like those acquired online and implemented in clinical contexts. After they finished, they were instructed to click a button to move on to the next one. Participants were not allowed to go back to previous pages to revise their recall answer. Participants completed the narrative task for all twenty videos in random order. The narrative recall task was self-paced and there was no time limit.

Data analysis

Hand-scored approach To score narrative recalls, we first generated an answer sheet for each video to score participants' narrative recalls against (Fig. 1). The answer sheet consisted of separating individual video clips into meaningful action segments, and each action segment represented one possible point on the answer sheet. The primary answer sheet was generated by one of the authors (X.S.). These scoring sheets were validated against answer sheets based on video descriptions from online participants. Twenty participants recruited from Prolific were asked to write step-by-step descriptions of the magic clips. Then a researcher (X.S.) combined magic clip descriptions from the 20 participants to generate a combined answer sheet to compare the validity of the original answer sheet. Specifically, all participants' descriptions of the videos were put together, with the overlapping descriptions removed and all different wordings left in the alternative answer sheet. For example, "the paper was cut" and "the paper was sliced" were combined as "the paper was cut and sliced." Intraclass correlation coefficients (ICCs; absolute agreement, two-way mixed effect) were compared using the two answer sheets. We found no difference in ICC scores using the original answer sheet and the combined answer sheet, suggesting that the original answer sheet was valid to use.

To score individuals' accuracy in recalling the videos, we first segmented each participant's narrative recalls into clauses. Clauses of participants' recall reflected individual action steps in the video clips, which was independent of the answer sheet. An example of a narrative recall was "The man takes a children's book full of pictures of candy, and then he shakes the book. The candy falls out of the book, after he reopens the book, all the candy pictures are gone." It was segmented into five clauses: "The man takes a children's book full of pictures of candy," "and then he shakes the book," "Then candy falls out of the book," "after he reopens the book," "all the candy pictures are gone." Next, two raters scored each clause of participants' recalls for each video independently by comparing the clause of participants' recall to each action step on the answer sheet. If a clause of recall matched an action step on the answer sheet, that clause of recall was scored as present by the rater (i.e., score = 1). Action steps on the answer sheet that did not have a

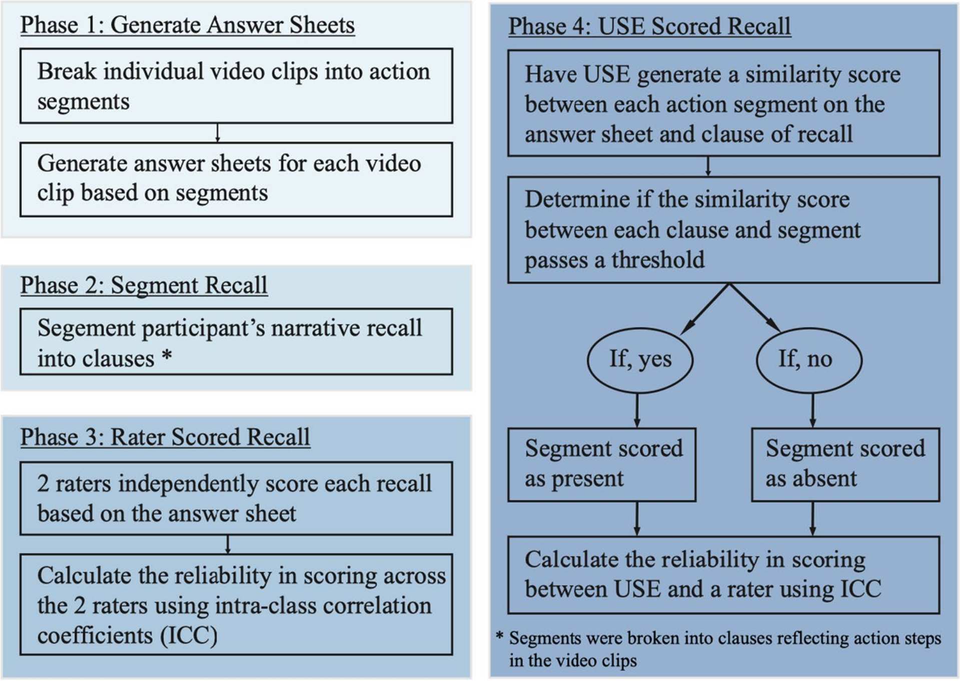
Phase 1: Generate Answer Sheets
Break individual video clips into action segments
Generate answer sheets for each video clip based on segments
Phase 2: Segment Recall
Segement participant's narrative recall into clauses *
Phase 3: Rater Scored Recall
2 raters independently score each recall based on the answer sheet
Calculate the reliability in scoring across the 2 raters using intra-class correlation coefficients (ICC)
Phase 4: USE Scored Recall
Have USE generate a similarity score between each action segment on the answer sheet and clause of recall
Determine if the similarity score between each clause and segment passes a threshold
If, yes
If, no
Segment scored as present
Segment scored as absent
Calculate the reliability in scoring between USE and a rater using ICC
* Segments were broken into clauses reflecting action steps in the video clips

Fig. 1 Flowchart of analysis with USE

corresponding clause were scored as absent by the rater (i.e., score = 0). The total score of a video was the sum of clauses present for that video. Finally, reliability in scoring across the two raters was calculated using intraclass correlation coefficients (ICC). ICC was used instead of a Pearson correlation because ICC represents a better measure of group homogeneity than the Pearson correlation (G. Chen et al., 2018). In both studies, ICC was calculated at both the clause level and video level, reflecting memory at varying levels of resolution. At the clause level, scoring of each video for each subject was compared between the two raters (or between individual raters and USE). At the video level, scoring of each video across all subjects was compared between the two raters (or between individual raters and USE).

Scoring with USE To automate the process of scoring narrative recalls, we applied the Universal Sentence Encoder (USE) to score participants' narrative recalls. The USE is a machine learning tool that converts text into high-dimensional vectors that can be used for semantic similarity analysis. Different from many machine learning tools that are trained and optimized for words, USE is trained and optimized for greater-than-word length text, such as sentences and phrases (Cer et al., 2018), which is optimal for analyzing sentence-based narrative recalls. USE was first used to convert clauses of participants' recalls and answer sheets into high-dimensional vectors. It then compared each clause of recall to each action step on the answer sheet and generated a similarity score using cosine similarity. Higher similarity score indicated higher similarity between the clause of recall and the action step on the answer sheet. Then, to determine when a clause would be scored as present, we need to determine a threshold used in the analysis. To find the optimal threshold, we tested threshold similarity score from 0.5 to 1.0, with step size of 0.1. Optimal threshold for Rater 1 and USE maximized ICC for Rater 1 and USE. To validate the optimal threshold, we applied optimal threshold for Rater 1 and USE to calculate ICC for Rater 2 and USE. Similarly, optimal threshold for Rater 2 and USE maximized ICC for Rater 2 and USE but was used to calculate ICC for Rater 1 and USE. The optimal threshold ranged from 0.53 to 0.77 for our data set. A clause of recall was scored as present if the similarity score between the clause and answer sheet was above the threshold. After all clauses of participants' recalls were scored, we calculated averaged ICC between ICC for Rater 1 and USE, and ICC for Rater 2 and USE. Same as the hand-scored approach, ICC was calculated at both the clause level and video level. Finally, to fully automate the process of scoring narrative recalls with USE, instead of using hand-segmented clauses by raters, we used an automated algorithm to auto-segment participants' recalls to clauses (automated approach) and calculated the reliability in scoring auto-segmented clauses between USE and raters. For the automated approach, we segmented participants' recalls using separators (i.e., but, then, or, however, which, that, and, ".", ",", ";"). The analysis was also run at both the clause level and video level. For those who would like to score narrative recalls automatically with USE, we recommend first scoring a subset of data manually, then calculating ICC between USE and raters to determine the optimal threshold before applying USE to the entire data set (follow Fig. 1 flowchart). Implementation of USE could be found online (https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder).

Scoring with GloVe To confirm the reliability of scoring narrative recalls with USE, we further tested the reliability of scoring with another machine learning tool, Global Vectors for Word Representation (GloVe; Pennington et al., 2014). Different from USE, GloVe converts individual word, rather than sentences, into vectors. We adapted codes from the following site: (https://github.com/mavalliani/Semantic-Similarity-of-Sentences). In the analysis, similar to scoring with USE, GloVe generated similarity scores by comparing clause of recalls to action steps on the answer sheet. We averaged the word embedding for each word in each clause to obtain a clause vector. We compared clause vectors using cosine similarity. Then, same as USE, we compared ICC between GloVe and individual raters using a threshold similarity score from 0.5 to 1.0 to find the optimal threshold. A clause was scored as present if the similarity score passed the threshold. The sum of the number of clauses present was the total score for a video. Finally, we calculated the reliability in scoring between GloVe and individual raters using ICC. Analyses were run for both hand-segmented clauses and auto-segmented clauses, and at both clause level and video level.

Comparison of hand-scored approach and automated approach To test whether USE is as reliable as human raters in scoring narrative recalls, four repeated analysis of variance (ANOVA) models were run in Python (Version 3.6.3) to compare reliability in scoring for hand-segmented clauses and for auto-segmented clauses at both clause level and video level. ICC for each participant was submitted as the dependent variable and scoring approach (raters, USE, and GloVe) was submitted as the independent variable. Bayes factors were calculated using ttestBF and anovaBF functions from the BayesFactor package in R (Version 4.1.1). Intraclass correlation coefficients and their 95% confidence interval were calculated using the intraclass_corr function from the pingouin package (Version 0.4.0) in Python based on an absolute agreement, two-way mixed-effect model. We also calculated cross-correlation as a direct comparison of different scoring approaches, which confirms our results regarding reliability of USE in scoring free recalls against the hand-scoring approach. Cross correlation coefficient was calculated using the corrcoef function from the numpy (Version 1.20.3) package in Python.

Results

Reliability of automated scoring procedures on course resolution narrative recall

We first compared the reliability of using automated approaches to score narrative recalls at the video level, which represents more coarse resolution details of memoranda. Across both our hand-scoring and automated methods, ICC scores between two raters, ICC scores between USE and raters, and ICC scores between GloVe and raters were excellent (Table 1), demonstrating that all methods were suitable for analysis of narrative recall data. To determine the reliability of each method against traditional hand-scoring approaches, we also utilized cross-correlation analyses. We found strong, significant cross correlation between the hand-scored approach and both USE-scored and GloVe-scored approach (Table 2). Notably, direct comparison between USE and GloVe showed no difference in reliability of scoring when narratives were hand-segmented ($t = -0.46$, $p = .65$, BF = 0.26).

To fully automate the process of scoring narrative recalls, we then developed an automated algorithm to auto-segment narrative recalls and re-tested the reliability across approaches. When recall was auto-segmented rather than hand-segmented, ICC scores between USE and raters remained good, while ICC scores between GloVe and raters were moderate. Cross-correlation between hand-scored approach and USE-scored approach remained significantly strong (Table 2), while the correlation between the hand-scored and GloVe-scored approach was moderate. Direct comparison between USE and GloVe showed that the USE-scored approach was significantly more reliable than the GloVe-scored approach ($t = 3.65$, $p < .01$, BF = 23.16). Overall, we showed that our automated segmentation approaches were suitable for reliably characterizing narrative recall data and reflected measures yielded by hand scoring. Additionally, we found that USE outperformed GloVe in scoring.

Reliability of automated scoring procedures on fine resolution narrative recall

We next tested the reliability of using automated approaches to score narrative recalls at a more fine-grained approach by comparing ICC scores across scoring methods at the clause level, which reflects each individual action in a movie clip as separate memoranda. At the clause level, ICC scores between two raters, ICC scores between USE and raters, and ICC scores between GloVe and raters were good (Table 1). Cross-correlation between the hand-scored approach and both the USE-scored approach and GloVe-scored approach were significant and strong (Table 2). There was no difference in scoring between USE and GloVe when narratives were hand-scored at the clause level ($t = 0.50$, $p = .63$, BF = 0.26).

Table 1 Summary of ICC reliability scores across hand scored and automated approaches

Study 1	ICC at clause level	ICC at video level
ICC between two raters	0.78, CI [0.73, 0.81]	0.93, CI [0.83, 0.97]
ICC between USE and raters	0.78	0.92
ICC between GloVe and raters	0.79	0.91
ICC between USE and raters: auto-segmented	0.74	0.86
ICC between GloVe and raters: auto-segmented	0.63	0.73

ICC assessments: Moderate = 0.5 < ICC < 0.75; Good = 0.75 < ICC < 0.9; Excellent = ICC > 0.9. *Note.* No confidence intervals were calculated for ICCs between raters and the automated approach because average ICCs were used in the table

Table 2 Summary of cross correlation coefficient across hand scored and automated approaches

	Cross correlation coefficient (clause level)	Cross correlation coefficient (video level)
Hand-scored approach and USE-scored approach	0.76, $p < .001$	0.83, $p < .001$
Hand-scored approach and GloVe-scored approach	0.76, $p < .001$	0.75, $p < .001$
Hand-scored approach and USE-scored approach: auto-segmented	0.78, $p < .001$	0.73, $p < .001$
Hand-scored approach and GloVe-scored approach: auto-segmented	0.55, $p < .001$	0.63, $p < .001$

Cross correlation coefficient: Moderate = 0.4 < r < .70; Strong = 0.70 < r < .90; Perfect = r = 1.0 (Akoglu, 2018)

When recall was auto-segmented rather than hand-segmented, ICC scores for USE and raters, and ICC scores for GloVe and raters were moderate. The cross correlation between hand-scored approach and USE-scored approach remained strong, but the correlation between the hand-scored approach and the GloVe-scored approach was only moderate. When comparing automated-segmentation approaches at the clause level, we found that USE-scored approach was significantly more reliable than the GloVE-scored approach ($t = 3.91$, $p < .01$, BF = 38.68). In sum, the same pattern of results that we found for the video-level was present at the clause level.

Study 2

Study 1 showed that our automated scoring approach provided reliable estimates of accuracy. However, this first sample was limited to a college-aged normative sample. To ensure that our automated segmentation approach could be widely useable across diverse populations, we tested the validity of the approach on populations that vary on clinically relevant dimensions known to influence episodic memory, specifically age and anxiety symptoms. We first confirmed that these variables (age, anxiety) influenced narrative recall accuracy using hand-scored approaches, and then examined the reliability of our automated approach in scoring data from these populations that show memory deficits. Given that in Study 1 we found that the USE-scored approach outperformed the GloVE-scored approach for auto-segmented recalls, in Study 2, we limited our analyses to the USE-scored approach.

Materials and methods

Participants

In Study 2, to further test the validity of USE in scoring large-scale individual difference data, 40 participants drawn from different age groups (young: 18–25; old: 65–79) and anxiety scores (low: 29–44; high: 65–114) were recruited from the Prolific online testing platform. Temple University's Institutional Review Board approved all study materials and procedures. All participants provided informed consent and were compensated for $6.5/hour.

Stimuli

Five high-curiosity video clips and five low-curiosity video clips were selected from the stimuli set from Study 1. Average curiosity rating for high-curiosity video stimuli was 6.40 (out of 10) and standard deviation for high-curiosity video stimuli was 0.49; average curiosity rating for low-curiosity stimuli was 4.74 (out of 10) and the standard deviation was 0.53. Length of magic clips were matched for high-curiosity stimuli and low-curiosity stimuli. The average length of high-curiosity stimuli was 30 seconds, and the average length of low-curiosity stimuli was 21 seconds. These videos were selected for providing the highest ICC scores between two raters in Study 1. High-curiosity videos were S32, K2, K12, S16, K4, and low-curiosity videos were K25, S30, S21, K24, S3_short.

Procedure

To measure participants' anxiety level, on day one, a total of 100 participants (50 young) completed two questionnaires regarding anxiety and depression (PROMIS Bank v1.0–Anxiety and PROMIS Bank v1.0–Depression) before watching the ten magic clips in random order. After a 24-hour delay, same as Study 1, participants were presented with screenshots of magic clips they watched on the first day in random order and were instructed to write in one text box a step-by-step recreation of the magic clips based on their memory of the screenshots. Of the 100 tested participants, we then selected 40 participants (20 young). Within each age group, half of the selected participants were of high anxiety level and half were of low anxiety level. The participants selection ensured significant differences in age and anxiety levels for different groups.

Data analysis

To test the effect of age on recall memory accuracy, a *t* test for memory accuracy using the hand-scored approach was performed between young population and old population in R (Version 4.1.1). To test whether our automated approach with USE was valid for performing large-scale individual difference analysis, we analyzed Study 2 data with both the hand-scored approach and automated approach with USE (see Study 1). A 2 × 2 (age group: old vs. young; approach: hand-scored vs. USE-scored) ANOVA was performed in Python (Version 3.6.3) statsmodel (Version 0.12.2) to test whether USE was reliable in scoring participant with varying ages. Another 2 × 2 (anxiety group: high vs. low; approach: hand-scored vs. USE-scored) ANOVA was also performed to test whether USE was reliable in scoring participants with varying anxiety levels.

Results

Narrative recall accuracy decreased with age and anxiety.

To confirm the effect of age on memory accuracy, we compared narrative recall accuracy between young and old population. We found that young population had greater accuracy than the old population in narrative recall ($t = 2.05$, $p = .04$) using the hand-scored approach, suggesting that memory decreased with age. We also compared narrative recall accuracy between the low-anxiety and high-anxiety population. We found that the low anxiety population outperformed the high-anxiety population ($t = 3.35$, $p < .01$), suggesting that memory decreased with anxiety.

Reliability of automated scoring procedures on coarse resolution narrative recall generalizes across age and anxiety groups

We then tested the reliability of USE in scoring these individual difference data. At the video level, we replicated our findings from Study 1, showing that across all participants, ICC between USE and raters were excellent (Table 3). To determine if our automated approach was generalizable across populations varying on clinically relevant variables, we compared whether the reliability of the hand-scored and USE-scored approach differed as a function of age and anxiety. Despite there being significant differences in free recall accuracy across groups, a general linear model did not show a main effect of group (age: $F = 0.74$, $p = .39$, BF = 0.30), or an Age × Approach interaction ($F = 0.67$, $p = .42$, BF = 0.13) on ICC scores, suggesting equal reliability in scoring free recall across age. Similarly for anxiety, despite there being significant differences in free recall accuracy for high anxiety participants, a general linear model did not show a main effect of group (anxiety: $F = 0.94$, $p = .33$, BF = 0.33), or an Anxiety × Approach interaction ($F = 0.002$, $p = .97$, BF = 0.10), suggesting equal reliability across varying anxiety symptoms. In sum, our results suggested that the USE-scoring approach remained valid for scoring individual difference data and did not introduce any systematic biases across populations.

Reliability of automated scoring procedures on fine resolution narrative recall generalize across age and anxiety groups

At the clause-level, we found that ICC scores between two raters and ICC scores between USE and raters remained between moderate and good across methods. There was no main effect of group or a Group × Approach interaction in individuals varying age (age: $F = 0.76$, $p = .39$, BF = 0.31; Age × Approach: $F = 0.56$, $p = .46$, BF = 0.12) or anxiety (anxiety: $F = 2.36$, $p = .13$, BF = 0.56; Anxiety × Approach: $F = 0.39$, $p = .53$, BF = 0.23). These findings suggest that our automated scoring approach was reliable in scoring large-scale individual difference data and would not introduce biases in across group analyses.

Table 3 Summary of ICC reliability scores across hand scored and automated approaches

Study 2	ICC at clause level	ICC at video level
Hand-scored approach	0.84	0.90
USE-scored: auto-segmented	0.67	0.88

ICC assessments: Moderate = 0.5 < ICC < 0.75; Good = 0.75 < ICC < 0.9; Excellent = ICC > 0.9

General discussion

To reliably score large-scale narrative recall and to reduce systematic human bias, we tested the reliability of using a natural language processing tool, the Universal Sentence Encoder (USE), to score narrative recalls. In Study 1, at the clause level, we found a moderate to good ICC score between hand-scored raters and our newly developed USE approach, depending on whether free recalls were automatically segmented or segmented by hand, respectively. Similarly, we found good to excellent ICC scores at the video level. There was also a significant and strong correlation between USE-scored and hand-scored approaches, supporting that USE is a promising tool to score large-scale narrative recalls. In Study 2, we further tested if USE remained valid for scoring individuals with varying ages and anxiety scores, which allows us to generalize the approach to clinically relevant populations. We found that even though the young population scored higher than the older population in the narrative recall task, the reliability of the scoring, as assessed by ICC between USE and hand scoring, was not different. Again, we found a moderate to good ICC score between raters and USE at both clause and video level. There was also a significant and strong cross correlation between the USE-scored and hand-scored approaches. In sum, the two studies showed that USE is a reliable tool to score large scale individual difference data.

ICC has been widely used in psychology and medical research to evaluate interrater, test–retest, and intrarater reliability (Koo & Li, 2016). Normally, ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability (Koo & Li, 2016). Prior research suggested that values of 0.60 or 0.70 are often used as the minimum standards for acceptable reliability coefficients, but this varies for studies with different research purposes (Shoukri et al., 2004; Terwee et al., 2007). In Study 1, we found a minimum ICC score of 0.74 and in Study 2, we found a minimum ICC score of 0.67 with USE (automated approach), suggesting that our automated approach with USE has yielded appropriate scores for general use.

Further, we found that these scores increased when individual narrative recall data was segmented into clauses by hand, rather than using our automated segmentation approach. To further confirm the reliability of our automated approach, we also examined the cross-correlation between the hand-scored approach and automated approach. We found that USE captured up to 75% of the variance of the hand-scored approach, suggesting that our automated approach strongly reflects the hand-scored approach. We used written recalls in the analyses, but given potential differences between written recalls and transcribed verbal recalls, further validation of the Universal Sentence Encoder with transcribed verbal recalls is needed to confirm the generalization of the USE-scored approach.

We also fully automated the process of narrative recall scoring by developing an automated algorithm with USE that automatically segmented narrative recalls into clauses and scored the segmented recalls. There was a significant and strong cross-correlation between the automated approach with USE and the hand-scored approach. More importantly, although USE and GloVe performed equally well in scoring hand-segmented recalls, we found that USE outperformed GloVE in scoring auto-segmented clauses, which is predicted because USE embeds not only words but phrases and sentences while GloVe only embeds words. Thus, we recommend using USE over more word-embedding methods, when utilizing an automated scoring approach to narrative recalls. This automated approach with USE could be used by other studies that use naturalistic stimuli.

Our studies showed that USE outperformed GloVe and was reliable in scoring large-scale individual difference data. However, ICC between raters and USE was still lower than ICC between two independent raters (the hand-scored approach), suggesting that the automated approach implementing USE has its own limitations compared to the hand-scored approach. Thus, there is a trade-off between accuracy and efficiency, such that the automated approach with USE was less accurate compared to the hand-scored approach, but provides an opportunity to run much larger samples, thus yielding more reliable findings. Notably, while NLP would avoid some of the systematic biases introduced by hand scoring, USE may include its own set of limitations from overemphasizing concrete nouns/language when determining similarity between retrieved memories and the actual event. However, it is difficult to assess which scoring procedure provides a more accurate description of memories for actual events. Future studies, however, could combine these two approaches to determine their utility in capturing individual differences across participants and engagement of relevant learning systems (i.e., hippocampus).

In Study 2, critically, we found that there were no systematic biases in this automated approach in capturing free recall accuracy when looking at clinically relevant populations with significant impairments in episodic memory, including older adults and individuals scoring high on self-reports in anxiety. In this way, these findings show that narrative recall paradigms can feasibly be utilized in large samples to assess individual differences in memory performance. Idiosyncratic to our data set, this approach can be utilized to assess how these populations may vary in mechanisms that embody curiosity-related memory enhancements (Gruber & Ranganath, 2019). We hope that these findings can be leveraged to introduce more nuanced forms of memory assessments in large-scale clinical studies. This approach could also be applied to other datasets to examine the effect of individual differences using naturalistic stimuli. Notably, this approach has not been validated with other naturalistic datasets. Future studies are needed to confirm the reliability of scoring narrative recalls with USE, in particular for stimuli that use longer narratives (i.e., Full TV Episodes, Movie Clips).

In summary, our two studies suggested that although the automated approach with USE is slightly less accurate than the hand-scored approach, scoring narrative recalls using the automated approach with USE is reliable and efficient, and no systematic biases are introduced when scoring individuals with varying ages and anxiety scores. The magic clips and the automated scoring approach with USE can be used together as a toolbox to assess the influence of curiosity on clinically relevant populations. The automated scoring approach can also be used more broadly in studies using naturalistic stimuli.

Acknowledgements We thank Eliza Meketon for assistance in data collection and analysis.

Author contributions X.S., D.V.S., and V.P.M. designed research; X.S., T.H. performed research; X.S. performed the analysis; and X.S, T.H., D.V.S., V.P.M wrote the paper.

Funding The authors did not receive support from any organization for the submitted work.

Data availability The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval Approval was obtained from the ethics committee of Temple University. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participants Informed consent was obtained from all individual participants included in the study.

Conflicts of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91–93.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder. *ArXiv:1803.11175 [Cs]*. http://arxiv.org/abs/1803.11175. Accessed 24 Nov 2021

Chen, J., Leong, Y. C., Norman, K. A., & Hasson, U. (2016). Shared experience, shared memory: A common structure for brain activity during naturalistic recall. *Neuroscience*. Advance online publication. https://doi.org/10.1101/035931

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience, 20*(1), 115–125.

Chen, G., Taylor, P. A., Haller, S. P., Kircanski, K., Stoddard, J., Pine, D. S., Leibenluft, E., Brotman, M. A., & Cox, R. W. (2018). Intraclass correlation: Improved modeling approaches and applications for neuroimaging. *Human Brain Mapping, 39*(3), 1187–1206.

Coutanche, M. N., Koch, G. E., & Paulus, J. P. (2020). Influences on memory for naturalistic visual episodes: Sleep, familiarity, and traits differentially affect forms of recall. *Learning & Memory, 27*(7), 284–291.

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

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience, 33*(48), 18906–18916.

Gruber, M., & Ranganath, C. (2019). How curiosity enhances hippocampus-dependent memory: The Prediction-Appraisal-Curiosity-Exploration (PACE) Framework. *Open Science Framework*. Advance online publication. https://doi.org/10.31219/osf.io/5v6nm

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Naspi, L., Hoffman, P., Devereux, B., & Morcom, A. (2021). Perceptual and semantic representations at encoding contribute to true and false recognition of objects. *Neuroscience*. Advance online publication. https://doi.org/10.1101/2021.03.31.437847

Ozono, H., Komiya, A., Kuratomi, K., Hatano, A., Fastrich, G. M., Raw, J., Haffey, A., Meliss, S., Lau, J. K. L., & Murayama, K. (2020). Magic curiosity arousing tricks (MagicCATs): A novel stimulus collection to induce epistemic emotions. *PsyArXiv*. https://doi.org/10.31234/osf.io/qxdsn

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Ren, Y., Nguyen, V. T., Sonkusare, S., Lv, J., Pang, T., Guo, L., Eickhoff, S. B., Breakspear, M., & Guo, C. C. (2018). Effective connectivity of the anterior hippocampus predicts recollection confidence during natural memory retrieval. *Nature*. *Communications, 9*(1), Article 4875.

Saarimäki, H. (2021). Naturalistic stimuli in affective neuroimaging: A review. *Frontiers in Human Neuroscience, 15*, Article 675068.

Scheurich, R., Palmer, C., Kaya, B., Agostino, C., & Sheldon, S. (2021). Evidence for a visual bias when recalling complex narratives. *PLOS ONE, 16*(4), Article e0249950.

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research, 13*(4), 251–271.

Silva, M., Baldassano, C., & Fuentemilla, L. (2019). Rapid memory reactivation at movie event boundaries promotes episodic encoding. *The Journal of Neuroscience, 39*(43), 8538–8548.

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences, 23*(8), 699–714.

St-Laurent, M., Moscovitch, M., & McAndrews, M. P. (2016). The retrieval of perceptual memory details depends on right hippocampal integrity and activation. *Cortex, 84*, 15–33.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42.

Wang, X., Xu, Y., Wang, Y., Zeng, Y., Zhang, J., Ling, Z., & Bi, Y. (2018). Representational similarity analysis reveals task-dependent semantic influence of the visual word form area. *Scientific Reports, 8*(1), Article 3047.

Zhu, Y., Yan, E., & Wang, F. (2017). Semantic relatedness and similarity of biomedical terms: Examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Medical Informatics and Decision Making, 17*(1), Article 95.

Open practices statement The data and materials will be available after publication upon request. None of the experiments was preregistered.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.