View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Munin - Open Research Archive

Psychiatry Research 297 (2021) 113743

Contents lists available at ScienceDirect

Psychiatry Research

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

Extending the usefulness of the verbal memory test: The promise of machine learning

Chelsea Chandler [a,b,*], Terje B. Holmlund [c], Peter W. Foltz [b,d], Alex S. Cohen [e], Brita Elvevåg [c,f,*]

[a] Department of Computer Science, University of Colorado Boulder, CO, USA
[b] Institute of Cognitive Science, University of Colorado Boulder, CO, USA
[c] Department of Clinical Medicine, University of Tromsø - The Arctic University of Norway, Norway
[d] Pearson, CO, USA.
[e] Department of Psychology, Louisiana State University, LA, USA.
[f] Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway

ARTICLE INFO

Keywords:
Machine learning
Natural language processing
Verbal memory test
Criteria

ABSTRACT

The evaluation of verbal memory is a core component of neuropsychological assessment in a wide range of clinical and research settings. Leveraging story recall to assay neurocognitive function could be made more useful if it were possible to administer frequently (i.e., would allow for the collection of more patient data over time) and automatically assess the recalls with machine learning methods. In the present study, we evaluated a novel story recall test with 24 parallel forms that was deployed using smart devices in 94 psychiatric inpatients and 80 nonpatient adults. Machine learning and vector-based natural language processing methods were employed to automate test scoring, and performance using these methods was evaluated in their incremental validity, criterion validity (i.e., convergence with trained human raters), and parallel forms reliability. Our results suggest moderate to high consistency across the parallel forms, high convergence with human raters (r values ~ 0.89), and high incremental validity for discriminating between groups. While much work remains, the present findings are critical for implementing an automated, neuropsychological test deployable using remote technologies across multiple and frequent administrations.

1. Introduction

Neuropsychological functioning is typically assessed in a professional setting during a dyadic exchange between a patient and a psychometrician. For this reason, neuropsychological assessment requires considerable resources on the part of the patient and the professional, and is not optimized for repeated administration within an individual over time (see McCaffrey and Westervelt, 1995; Ruff, 2003). Recent advances in technology are making it possible to monitor psychiatric states frequently in a remote manner, transforming how health information is generated (Ben-Zeev et al., 2015; Torous and Baker, 2016; Holmlund et al., 2019a; Chandler et al., 2019; Cohen et al. 2020a; for a review see Tal & Torous, 2017). The present study is part of a larger program to increase the value of monitoring spoken communication consensually through digital channels (Cohen et al., 2019; Holmlund et al., 2019a; Chandler et al., 2020c). The goal of this work is to objectively quantify speech with natural language processing (NLP) and machine learning techniques in order to provide accurate indicators of cognitive and mental health with applications in neurology, psychiatry, and behavioral assessment (Cohen et al., 2019; Holmlund et al., 2019a; Chandler et al., 2020c). The present project examined the use of a novel verbal memory test (story recall) that enables automated, remote, and frequent administration.

1.2. An evolving field

Verbal episodic memory has traditionally been assessed by participants learning prose passages and then subsequently recalling the stories. Scores are given by counting the number of units of information recollected as in the popular Logical Memory prose recall task of the Wechsler Memory test (now in its seventh decade and fourth revision - Wechsler, 1945, 1987, 1997, 2009). These story units must be recollected with specific predefined verbiage and is scored using a 25 point rubric, with a single point given to each story unit recalled verbatim.

* Corresponding authors.
E-mail addresses: chelsea.chandler@colorado.edu (C. Chandler), brita.elvevag@uit.no (B. Elvevåg).

https://doi.org/10.1016/j.psychres.2021.113743
Received 23 June 2020; Accepted 16 January 2021
Available online 19 January 2021
0165-1781/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Many times patients recall a story with small additions, deletions, and substitutions, while still recollecting the gist. More recent versions of the Wechsler Memory test account for thematic units where patient recalls are judged for accurately recalling the theme, in addition to recalling exact story units. This measure allows for points to be given to the recalls of patients who correctly remember the gist of the story, but use the wrong verbiage. However, Dunn et al. (2002) contend that the thematic units add no additional information as they are merely a subset of the original story units.

Computational approaches to model language features in story recall tasks allow us to move beyond simple counting of the amount of items recalled and enable a more nuanced approach to the analysis of what is actually recalled (Rosenstein et al., 2014; Lautenschlager et al., 2006). NLP allows the language spoken by patients to be analyzed in multiple ways: at a word-level, structurally, and semantically. Word level characteristics of language include measures such as simple counts of words, parts of speech, phrases, and words related to cognitive and affective processes (Pennebaker et al., 2015; Prud'hommeaux and Roark, 2011). Structural characteristics, such as n-grams, sentence parses, and cohesion, measure how well language is "put together", allowing for characterization of syntactic structures and flow fluency of expression. Semantic characteristics, often modeled using word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2018; Brown et al., 2020), are able to encode the underlying meaning of words, sentences, or whole passages in order to judge the appropriateness of the meaning expressed. Furthermore, mobile technology affords for remote data collection, self-administration, and immediate analysis of the patient responses (Chandler et al., 2019; Holmlund et al., 2020; Chandler et al., 2020c).

Our approach to memory assessment is not simply digitalization of current methods (as is standard practice in the psychometric assessment industry), but allows for a completely novel method of assessment delivery and scoring. This radically different approach affords the potential for a full response processing pipeline that in principle enables more sensitive assessment (i.e., a more comprehensive and full-spectrum analysis of responses allows for a nuanced approach to modeling speech, semantics, and syntax). Such a framework starts with data collection on a mobile platform outside of the traditional laboratory or clinical setting, then to automated speech recognition transcription of patient speech, and finally to automated ratings of task completion. With such a framework, and as normative databases grow, these techniques will enable us to address issues of specificity. Furthermore, they will allow for more regular monitoring than current methods which, because there are a limited number of stories, can be maximally administered a couple of times a year, and enable automated remote assessments (Chandler et al., 2020c) thus potentially enhancing access and equity in healthcare. The present project examined preliminary psychometric characteristics of this test, administered to separate patient and nonpatient populations.

2. Methods

2.1. Participants

For the purpose of automatically assessing verbal memory in a population of 94 stable patients with mental illness recruited from an inpatient substance use treatment program, as well as 80 presumed healthy nonpatients recruited from a university, 24 verbal memory stories were generated. The patients had various diagnoses: substance abuse disorder (N = 34), depression (N = 31), anxiety disorder (N = 12), mood disorder (N = 6), bipolar disorder (N = 5), schizophrenia (N = 3), and post-traumatic stress disorder (N = 3). The mean age of the presumed healthy nonpatients was 19.85 (std = 1.96, min = 18, max = 27) and for the patients was 37.18 (std = 10.72, min = 19, max = 69). Nineteen of the presumed healthy nonpatients were male and 61 were female while all of the patients were male. Of the presumed healthy nonpatients, 59 were Caucasian, 11 were African-American, 4 were Asian-American, 4 were multi-racial, and 2 were other: not American. Of the patients, 51 were African-American, 39 were Caucasian, 3 were listed as N/A, and 1 was American Indian.

2.2. Story creation and rating

Each story in this study was developed to be structurally similar to the Logical Memory subtest of the Wechsler Memory Scale-R and Wechsler Memory Scale-III, which currently have two alternate test forms each, with one being nearly identical between the two sets (WMS-R, Wechsler, 1987; WMS-III, Wechsler, 2009). Thus in our analyses, we conflate the two nearly identical versions and compare our own story versions with the three WMS variations. Variations of our verbal memory tasks were presented over multiple sessions to the same participant in a spoken format via a smart device application developed to remotely administer neuropsychological tasks to its users (the *delta* Mental State Examination or *d*MSE; see Holmlund et al., 2019a and Chandler et al., 2020c for more details of the *d*MSE application). NLP techniques were applied to the recalls (both immediate recalls and recalls prompted a day after administration) to automatically generate predictions of expert human assessments.

The stories we created contained two characters, a setting, a problem, and a resolution. Our stories ranged from 68 to 82 words in length, whereas the original WMS stories ranged from 65 to 86 words. This property, as well as additional properties of the stories such as average words per sentence and number of sentences, as compared to the three WMS stories, are depicted graphically in Fig. 1 (see Appendix A for the data in table format with other properties). It is shown that the three WMS stories have a wider range in terms of their structural properties. The original WMS stories had a larger variation in the number of words used as compared to our original stories and notably had longer sentences. Furthermore, the original WMS stories tend to be Americentric, naming American cities and colloquialisms from American English. The creation of less ethnocentric stories will allow for the task to be more generalizable and thus be used in various locations with diverse populations.

The WMS story recalls are traditionally scored on a 25 point scale, with one point given for each predefined story unit recalled. Of importance is the requirement that the participants must recall these story units with exact or nearly exact verbiage. Scoring is a nuanced and time consuming process for clinicians. While this traditional method of scoring can be easily digitized by machines with simple pattern matching between rubrics and transcribed speech, this process is nevertheless unable to capture the continuous nature of semantics. Thus, we decided to move away from the 25 point exact matching rubric and instead created a simple 6 point rubric for scoring story recalls. A score of 1 indicated no details were recalled, and 6 indicated all major and almost all minor concepts and themes were recalled. Three trained human raters with clinical experience assigned scores on the general quality and amount of concepts and themes recalled, including characters, events, dates, descriptors, and feelings. Of the total 823 responses collected, a subset (N = 326) was rated by two raters to verify inter-rater reliability (r = 0.87). This high agreement suggests that the rubric was reliable and thus appropriate for use in training a machine learning model.

2.3. Natural language processing techniques

The typical process of counting overlapping story units can be trivially automated with pattern matching. This is the use of regular expressions, or sequences of characters, that define a search pattern with which we compare to a text passage to find matching segments (e.g., searching the string "Hello, world" for the pattern "He" will result in a match on the first word, but searching it for "and" will result in no matches). We propose that enriching this technique with vector-based

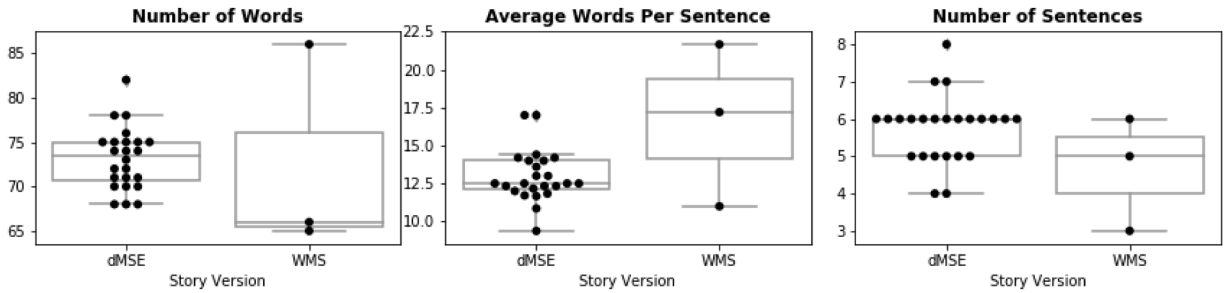

Fig. 1. Box plots of three structural properties of the *d*MSE story variations, as compared to the WMS story variations. From left to right, we show the number of words per story, the average number of words per sentence per story, and the number of sentences per story. Exact numbers and more properties are given in Appendix A.

NLP methods will allow for more fine-grained and nuanced assays of remembrance scores.

In the scoring of the verbal memory test, we have previously proposed the use of machine learning regression models trained on (1) the number of unique words spoken, (2) the number of common words between the original story and the recall, and (3) the word mover's distance between the original story and the recall (Chandler et al. 2019; Holmlund et al. 2020). The vector-based word mover's distance was our most highly correlated feature. The metric generates a mapping from each word embedding (a vector representation of a particular word; further discussed in the next paragraph) in one document to its closest counterpart in another document and the resulting distance is calculated as the sum of all Euclidean distances between matched word embeddings. When used in a machine learning regression model, these three features correlate with expert raters' 6 point rating scale with a Pearson r of 0.88. Thus, we have previously concluded that the verbal memory test can be sufficiently automated and that it holds much potential for real world applications (Chandler et al. 2019; Holmlund et al. 2020).

The use of word embeddings is especially important here because they are continuous vector space representations of language, calculated by the likelihood of certain words appearing close to one another in typical language. Typical language can be defined as the language found in Wikipedia, on the internet, or other large corpora such as books or movie transcripts. Semantics is *continuous* (Turney and Pantel, 2010) and vector-based analyses of language are able to capture the degree to which certain words are similar to others. The verbal memory test requires participants to remember salient words from the original story in their recall, yet this sort of task is rarely performed verbatim. Unfortunately, the traditional rating rubric that includes 25 binary yes/no units to count overlapping words does not capture the continuous nature of language. To further illustrate this point, we introduce a new feature for our regression models that alone is correlated r = 0.86 with expert ratings of immediate and delayed recalls: BERTScore (Zhang et al., 2020). The transformer-based contextual word embedding model, BERT (Devlin et al., 2018), has resulted in major improvements in many state of the art approaches to NLP tasks. BERTScore is an adaptation of BERT that is trained to compute an evaluation metric for text generation. Similar to other approaches, the metric computes a similarity measure of each token in a candidate sentence with one in a reference sentence, greedily chosen as the one with the highest similarity score. BERTScore is a relevant feature for story recall scoring because it adapts a state of the art NLP tool to compute the most accurate and*continuous* assay of similarity between an original story and its recall.

Specifically, the NLP implementation of comparing word embedding representations computes a cosine *similarity* between two word embedding vectors by subtracting the cosine distance between the two vectors from 1. In other words, two embeddings that are close in semantic space (i.e., similar or synonymous words) will have an angle or cosine *distance* close to 0, so their cosine *similarity* will be close to 1 since 1 minus ~0 is ~1. Conversely, the further the words are in semantic space (i.e., less similar), the closer their cosine *similarity* is to 0. As compared to this NLP implementation, where cosine similarity is on a continuous scale of 0 to 1, a binary manner of point allocation will not provide such a precise metric of amount recalled.

The following is an example of recall scoring in the traditional manner versus the vector-based NLP approach. An original variation of the WMS stories depicts a woman who is "employed as a cook". The traditional scoring rubric states that the recall must state this phrase verbatim or use any variation of the word "cook", such as "the woman cooked", but no points are given if the response mentions that the woman "is a chef", "works in a kitchen", or any other similar, yet non-verbatim, variation. Likewise, she works in a "cafeteria", yet the scoring rubric does not allow for synonyms such as "dining hall" or "lunchroom". Reciting "dining hall" in place of "cafeteria", for example, is less correct as it is not verbatim, but not entirely incorrect and this should be captured in the scoring metric. Fig. 2 shows how the cosine similarity between word embedding vectors captures semantics in a more precise manner as opposed to exact matching. We first show how BERT is able to capture that "works in a kitchen" and "is a chef" are fairly synonymous with "employed as a cook", but that "works as a janitor" is less so (which is the desired result as it is not synonymous). The traditional rubric would give all three of these variations a score of 0 even though their level of similarity is more continuous than a score of 0 would entail. We also show that BERT captures how extremely similar "dining hall" and "lunchroom" are to "cafeteria", and how "garage" is less similar, yet the traditional rubric would again assign 0 points to each of these phrases.

2.4. Analytic approach

Chapman and Chapman (1973) discussed the importance of discriminability when comparing scores from different classic psychometric tests. When comparing tests of the same format and mode of scoring, such as parallel forms of the verbal memory test, the Chapmans defined the discriminating power of a test as a function of average, spread, and covariance of item difficulty, as well as number of items. Furthermore, validation techniques such as exactly matching linguistic characteristics, testing on thousands of people, and controlling for representative participants are traditionally used (Schnabel, 2012). With computational approaches this is no longer critical, but there must be minimum requirements met for these stories to be viable. Thus, we take these approaches further with the inclusion of additional features. Whereas our previous work (Chandler et al. 2019; Holmlund et al. 2020) showed a principled way to automatically assess verbal memory, we now discuss the lessons learned concerning the suitability of each generated test variation in an automated clinical assessment setting. We evaluate three aspects of this test: 1) incremental validity (Section 3.1) - the degree to which multiple administrations provide improved explanatory power for differentiating between groups, 2) criterion validity (Section 3.2) - the degree to which our machine learning model predictions converge with expert human judgement of amount recalled across both patients and nonpatients, and 3) parallel forms reliability (Section 3.3) - the

Fig. 2. Graphical representation of two word embedding comparison settings. The first, on the left, is comparing "employed as a cook" to "works in a kitchen", "is a chef", and "works as a janitor". The BERT cosine similarities are 0.867, 0.862, and 0.683, respectively. The second, on the right, is comparing "cafeteria", with "dining hall", "lunchroom", and "garage". The BERT cosine similarities are 0.959, 0.956, and 0.886, respectively. Both scenarios portray a continuous spectrum of semantic similarity. The traditional WMS scoring rubric, on the other hand, would give all three variations a score of 0 points in each scenario.

degree to which reliable predictions of performance quartile are consistent over testing time of nonpatient individuals.

3. Results

3.1. Incremental validity: frequent administration entails a more accurate patient representation

We built a classification model of group membership (patient vs nonpatient) that was based on the same NLP features as our human rating regression model (see Chandler et al., 2019 for more classification model details). We limit our dataset to the 131 participants (53 patients and 78 nonpatients) who provided 4 or more responses to 4 or more story variations (i.e., interacted with the *d*MSE application on 4 or more separate occasions). We deemed this manner of selecting data as the most optimal amount to get the best possible ranking; there were not too few data points that the results would not show a trend and likewise there were not too many that we would miss a large subset of our data with such a restriction. In Table 1, we show that when starting with models trained and tested on 4 response samples from each participant, the overall classification accuracy is 83.2% (AUC = 0.857), and that each time a single response is removed from a participant's own subsample, the classification accuracy decreases, with a final overall accuracy of 74.1% (AUC = 0.782) when considering a single response. As might be expected, the results show that measurements from psychometric testing are more likely to reflect the underlying cognition that we are interested in when performance over multiple parallel forms of a task over time are considered. Furthermore, because of this criteria, in previous work, all classification was done by aggregating all responses over time of a single person rather than classifying based on one response only.

Table 1
Overall classification accuracy and area under the curve (AUC), patient accuracy (i.e., *sensitivity:* the ability of the model to correctly classify a patient from the pool of actual patients), and nonpatient accuracy (i.e., *specificity:* the ability of the model to correctly classify a person as nonpatient from the pool of nonpatient participants) when the model is trained and tested on 1 response from a person, 2 responses from a person, 3 responses from a person, and 4 responses from a person. As more recall responses are added to the classification model for a single participant, the ability to accurately classify each participant improves.

# stories administered	AUC	Overall accuracy	Patient accuracy	Nonpatient accuracy
1	0.782	74.1%	67.9%	78.2%
2	0.832	77.8%	75.5%	79.5%
3	0.845	80.1%	79.2%	80.8%
4	0.857	83.2%	81.1%	84.6%

3.2. Criterion validity: high correlation of model predictions to expert human ratings

Accurate automation of psychometric testing is a minimum requirement for the use of machine learning methods. In previous work, we were able to predict the rating an expert human would assign to participants' recalls by building a logistic regression model using NLP features (Chandler et al., 2019; Holmlund et al., 2020). Updating the previous models (as described in Section 2.3) with a new feature: BERTScore, in place of the word mover's distance, we have now found an average Pearson r correlation of 0.89 with expert human ratings, when performing a 10 fold cross-validation[1] through recalls of all stories. We found that all story variations performed well with human rating prediction in a range of r = 0.82-0.93, which is in line with the inter-rater correlation (r = 0.87) in this dataset. Such strong correlations with expert ratings are now possible more than ever before with the availability of more robust and generalizable word embedding models. Previous approaches to modeling language such as latent semantic analysis (Landauer et al., 1998) tend to be less universal as they were generally trained on smaller datasets with different training objectives (such as a matrix factorization in the case of latent semantic analysis) and therefore tend to underperform the more recent higher dimensional, deep neural network based approaches. The current state of the art in machine learning and natural language processing is now sufficiently advanced for the realization of these types of models.

3.3. Parallel forms reliability: reliable predictions in nonpatient individuals

We evaluate parallel reliability as the following: if an individual's performance on the verbal memory test is in the upper quartile of performance based on all of their *other* recalls, then they will also be in the upper quartile for the current variation of the task. The same would hold for each quartile. Thus, we expect an upward trend in a plot of performance against quartile. We chose quartiles as our assessment as this categorical approach could specifically support clinical decision making. In many cases it is possible to use the intraclass correlation, but the method we illustrate here helps visualize how the performance may differ across multiple forms of the assessment.

We performed this experiment on each story variation and found consistency in some stories but not others. For instance, in Fig. 3, plots a) and b) correspond to the results of a particular story, the CLOSED

[1] Cross validation is a method of splitting a dataset into partitions to train and test on in order to yield the most representative results over the samples. When the dataset is split into 10 equal subsamples, it is trained on nine and tested on the remaining single subsample. This process is done with each of the 10 splits serving as the test set in separate iterations, and then results from the 10 tests are averaged.

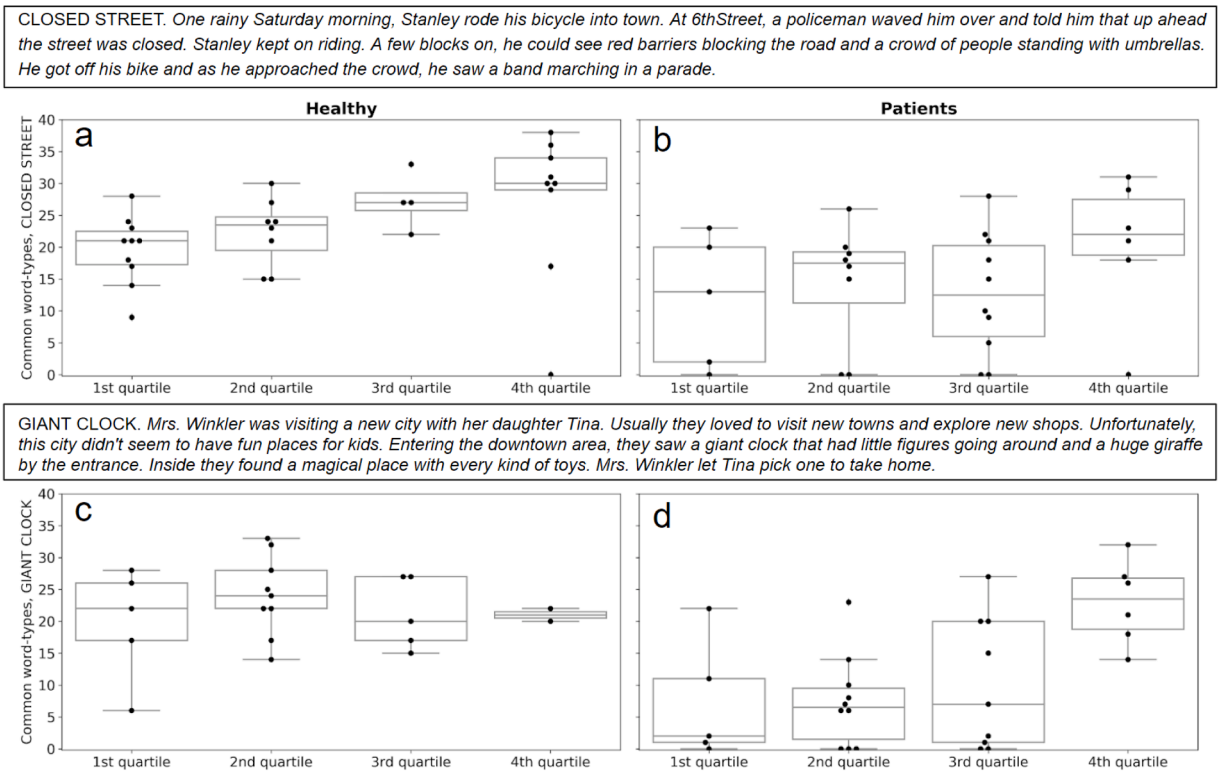

Fig. 3. Stories can differ in their ability to separate those who perform well in general from those who do not. ParticiSpants were ranked on the x-axis according to their overall performance on *other* story forms. The y-axis represents how well they performed on the CLOSED STREET (a, b) and GIANT CLOCK stories (c, d), represented by a count of common word-types between the retelling of a story and its original prompt. *Panel a*: On this story, nonpatient participants (N=33) showed the expected pattern where high performers had high scores. *Panel b:* In patients (N=29), the difference between high and low performers was not as obvious, and scores were generally lower. *Panel c:* The GIANT CLOCK story did not reveal differences between high and low performing nonpatient participants (N=21). *Panel d:* In patients (N=32), only the top performers were reliable, while the lower quartiles showed a floor-effect.

STREET story (see Fig. 3 for actual story), that showed reliable clustering. Plots c) and d), on the other hand, correspond to the GIANT CLOCK story (see Fig. 3 for actual story) which produced a downward trend in the quartile plot in the nonpatient population, and thus less reliability (e.g., performance of both patients and nonpatients is highly variable). An additional observation from these plots is that the performance of the patient population tended to show the floor effect (i.e., the measurements generally resulted in consistently low scores), which shows why we base story quality on the performance of a nonpatient population.

An intuitive reason for the CLOSED STREET story being more reliable than the GIANT CLOCK story could be because it is more relatable, whether it be from personal experience or in literature, media, and so on. The themes of *riding a bike* and *road blocks for parades* commonly occur in day-to-day life or participants will generally be familiar with the concept. The GIANT CLOCK story on the other hand is perhaps less relatable, especially given the confusion over what could be interpreted as entering a giant clock.

4. Discussion

In past work, we have argued that machine learning approaches must be explainable, transparent, and generalizable in order to be viable in a clinical setting (Chandler et al., 2020b). We also showed in a proof of concept study that neuropsychological tasks can be remotely administered via a smart device, that good quality data can be collected from patient application usage, and that various mental state variables can be predicted (beyond the straight-forward scoring of tasks) using machine learning methods (Chandler et al., 2020c). Now, we emphasize that it is essential that researchers clearly define which tests can be used in automation in order to produce sufficiently accurate results. This is a necessary step to move beyond the proof of concept research stage and translate into tools that generate actionable clinical inferences for patients.

Traditional psychometrics are, for the most part, based upon an era where behavioral responses by a participant were measured by an experimenter who summarized the performance by note taking, checking boxes on questionnaires and rubrics, or making a note of the response time, but this is rapidly changing as new methods for data collection and analysis are emerging. Currently, measurements of symptoms and signs of psychiatric disorders are often conflated with the underlying disorders themselves, but these limits must be carefully avoided in diagnostic assessment (Kendler, 2016). Put differently, when clinicians evaluate only the characteristics present in individual scoring manuals or, more generally, the items in the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2013), this leads to an impoverished view of psychopathology. With the reinvention of traditional diagnostics and monitoring to be suitable for automation, we must also reinvent the manner in which measurements are gathered and analyzed. It is critical that these new automated metrics are sensitive enough to enable the measurement of change over time, which necessitates new psychometrics (Cohen et al., 2020b). Furthermore, as an approach that has humans less 'in the loop', we must ensure that these new psychometric tests are not only at least of the same quality as the traditional human judgment-based testing, but that they could potentially move the field beyond this to detect subtleties that humans may miss (Topol, 2019; Ardila et al., 2019; Poplin et al., 2018; Grove et al., 2000).

Moving forward, there are numerous other factors that are necessary to account for such as collecting data from large samples of the

population across the lifespan, clinical conditions, cultural settings, and gender as well as recruiting diverse subgroups for usability testing. Such an approach to data collection will allow us to control for differences in demographic characteristics that have been shown to manifest in many applications of artificial intelligence in psychiatry (Cirillo et al, 2020). Specifically, we propose more frequent micro-level measurements (e.g., daily verbal memory tests where multiple modes of data are collected) rather than sparse macro-level measurements (e.g., an irregularly scheduled battery of psychological tests). This means consistent testing in small, unobtrusive ways, rather than a large slew of testing done on a greater time scale. In the age of big data, many modeling techniques are well-suited for this approach. Additionally, multi-modal observations tend to produce more accurate predictions of psychiatric variables of interest than those of uni-modal nature (Chandler et al., 2020a). Thus, we propose testing that is able to generate data of various modalities (e.g., vocalization features, language features, touch features, speed features, and so on; Holmlund et al. 2019b). Such a fine-grained approach will establish the necessary population norms to then make incisive inferences regarding the specific effects of various clinical parameters such as the effects of illness state, severity, medication, and neurobiological basis of the presenting or suspected underlying disorder. Naturally, this new approach to data collection, task design, and data analysis will require a clear road map as to how the new psychometrics will be validated, normed, and implemented. This paper offers a first step in this direction.

CRediT authorship contribution statement

Chelsea Chandler: Conceptualization, Methodology, Software, Writing - original draft, Visualization. **Terje B. Holmlund:** Conceptualization, Methodology, Software, Writing - original draft, Visualization. **Peter W. Foltz:** Conceptualization, Methodology, Writing - original draft. **Alex S. Cohen:** Conceptualization, Methodology, Writing - original draft. **Brita Elvevåg:** Conceptualization, Methodology, Writing - original draft.

Declaration of Competing Interest

None

Acknowledgements

Thank you to Miranda Lee-Foltz and Mark Rosenstein for the development of stories and to Miranda Lee-Foltz for rating of recalls. This project was funded by grant 231395 from the Research Council of Norway awarded to Brita Elvevåg

Appendix A. Table of structural properties of each of the 24 *d*MSE story variations and the 3 WMS story variations

Story	N words	N characters	N sentences	Words per sentence	Characters per word
*d*MSE-1	68	268	4	**μ**=17.00, **σ**=2.55	μ=3.94, **σ**=1.94
*d*MSE-2	75	336	6	**μ**=12.50, **σ**=5.06	**μ**=4.48, **σ**=1.82
*d*MSE-3	71	302	6	**μ**=11.83, **σ**=3.63	**μ**=4.25, **σ**=2.23
*d*MSE-4	73	333	6	**μ**=12.17, **σ**=6.44	**μ**=4.56, **σ**=1.97
*d*MSE-5	74	329	6	**μ**=12.33, **σ**=3.64	**μ**=4.45, **σ**=1.84
*d*MSE-6	70	296	5	**μ**=14.00, **σ**=6.10	**μ**=4.23, **σ**=2.27
*d*MSE-7	75	334	6	**μ**=12.50, **σ**=3.60	**μ**=4.45, **σ**=2.37
*d*MSE-8	70	271	6	**μ**=11.67, **σ**=2.21	**μ**=3.87, **σ**=1.79
*d*MSE-9	78	320	6	**μ**=13.00, **σ**=3.27	**μ**=4.10, **σ**=1.89
*d*MSE-10	76	372	7	**μ**=10.86, **σ**=4.79	**μ**=4.89, **σ**=2.33
*d*MSE-11	75	331	8	**μ**=9.38, **σ**=6.22	**μ**=4.41, **σ**=2.05
*d*MSE-12	74	330	6	**μ**=12.33, **σ**=4.23	**μ**=4.46, **σ**=2.09
*d*MSE-13	68	310	5	**μ**=13.60, **σ**=2.58	**μ**=4.56, **σ**=2.07
*d*MSE-14	75	331	6	**μ**=12.50, **σ**=2.63	**μ**=4.41, **σ**=2.00
*d*MSE-15	75	326	6	**μ**=12.50, **σ**=1.38	**μ**=4.35, **σ**=2.04
*d*MSE-16	71	301	5	**μ**=14.20, **σ**=3.06	**μ**=4.24, **σ**=1.93
*d*MSE-17	68	289	4	**μ**=17.00, **σ**=3.00	**μ**=4.25, **σ**=1.82
*d*MSE-18	70	292	5	**μ**=14.00, **σ**=6.20	**μ**=4.17, **σ**=1.93
*d*MSE-19	78	263	6	**μ**=13.00, **σ**=5.20	**μ**=3.37, **σ**=1.65
*d*MSE-20	72	308	5	**μ**=14.40, **σ**=7.34	**μ**=4.28, **σ**=2.06
*d*MSE-21	72	298	6	**μ**=12.00, **σ**=4.65	**μ**=4.14, **σ**=1.89
*d*MSE-22	74	285	6	**μ**=12.33, **σ**=4.50	**μ**=3.85, **σ**=1.74
*d*MSE-23	82	369	7	**μ**=11.71, **σ**=4.16	**μ**=4.5, **σ**=2.03
*d*MSE-24	71	279	5	**μ**=14.20, **σ**=2.04	**μ**=3.93, **σ**=1.89
WMS III/R-1	65	278	3	**μ**=21.67, **σ**=9.56	**μ**=4.28, **σ**=2.14
WMS III-2	86	371	5	**μ**=17.20, **σ**=7.78	**μ**=4.31, **σ**=2.40
WMS R-2	66	305	6	**μ**=11.00, **σ**=6.11	**μ**=4.62, **σ**=2.26

References

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, Fifth ed. American Psychiatric Publishing, Arlington, VA.

Ardila, D., Kiraly, A.P., Bharadwaj, S, et al., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. 25, 954–961 doi:0.1038/s41591-019-0447-x.

Ben-Zeev, D., Scherer, E.A., Wang, R., Xie, H., Campbell, A.T., 2015. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr. Rehabil. J. 38 (3), 218–226. https://doi.org/10.1037/prj0000130.

Brown, T.B., Mann, B., Ryder, N. et al. 2020. Language models are few-shot learners. *ArXiv* abs/2005.14165.

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: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pp. 137–147.

Chandler, C., Foltz, P.W., Cheng, J., Cohen, A.S., Holmlund, T.B., Elvevåg, B., 2020a. 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.

Chandler, C., Foltz, P.W., Elvevåg, B., 2020b. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. Schizophr. Bull. 46 (1), 11–14. https://doi.org/10.1093/schbul/sbz105.

Chandler, C., Foltz, P.W., Cohen, A.S., Holmlund, T.B., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Elvevåg, B., 2020c. Machine Learning for Ambulatory Applications

of Neuropsychological Testing, Volumes 1–2. Intelligence-Based Medicine. https://doi.org/10.1016/j.ibmed.2020.100006.
Chapman, L.J., Chapman, J.P., 1973. Problems in the measurement of cognitive deficits. Psychol. Bull. 79 (6), 380.
Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M.J., Chadha, A.S., Mavridiset, N., 2020. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digit. Med. 3, 81. https://doi.org/10.1038/s41746-020-0288-5.
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. J. Abnorm. Psychol. 128, 97–105. https://doi.org/10.1037/abn0000397.
Cohen, A.S., Cowan, T., Le, T.P., Schwartz, E.K., Kirkpatrick, B., Raugh, I.M., Chapman, H.C., Strauss, G.P., 2020a. Ambulatory digital phenotyping of blunted affect and alogia using objective facial and vocal analysis: proof of concept. Schizophr. Res. https://doi.org/10.1016/j.schres.2020.03.043.
Cohen, A.S., Cox, C., Tucker, R., Mitchell, K.R., Schwartz, E.K., Le, T., Foltz, P.W., Holmlund, T.B., Elvevåg, B., 2020b. Validating digital phenotyping technologies for clinical use: the critical importance of "resolution. World Psychiatry 19 (1), 114–115.
Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.
Dunn, J.C., Almeida, O.P., Barclay, L., Waterreus, A., Flicker, L., 2002. Latent semantic analysis: a new method to measure prose recall. J. Clin. Exp. Neuropsychol. 24, 26–35.
Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., Nelson, C., 2000. Clinical versus mechanical prediction: a meta-analysis. Psychol. Assess. 12, 19–30.
Holmlund, T.B., Foltz, P.W., Cohen, A.S., Johansen, H.D., Sigurdsen, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E., Elvevåg, B., 2019a. Moving psychological assessment out of the controlled laboratory setting and into the hands of the individual: Practical challenges. Psychol. Assess. 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 Res. 273, 767–769. https://doi.org/10.1016/j.psychres.2019.02.014.
Holmlund, T.B., Chandler, C., Foltz, P.W., Cohen, A., 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.
Kendler, K., 2016. The phenomenology of major depression and the representativeness and nature of DSM criteria. Am. J. Psychiatry 173. https://doi.org/10.1176/appi.ajp.2016.15121509 appi.ajp.2016.1.
Landauer, T.K., Foltz, P.W., Laham, D., 1998. Introduction to latent semantic analysis. Discourse Processes 25, 259–284.
Lautenschlager, N.T., Dunn, J.C., Bonney, K., Flicker, L., Almeida, O.P., 2006. Latent semantic analysis: an improved method to measure cognitive performance in subjects of non- English speaking background. J. Clin. Exp. Neuropsychol. 28 (8), 1381–1387.
McCaffrey, R.J., Westervelt, H.J., 1995. Issues associated with repeated neuropsychological assessments. Neuropsychol. Rev. 5 (3), 203–221.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR.
Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K., 2015. The Development and Psychometric Properties of LIWC2015. University of Texas at Austin, Austin, TX.
Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Empirical Methods in Natural Language Processing, pp. 1532–1543.
Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep Contextualized Word Representations. NAACL-HLT.
Poplin, R., Varadarajan, A.V., Blumer, K., et al., 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. 2, 158–164. https://doi.org/10.1038/s41551-018-0195-0.
Prud'hommeaux, E.T., Roark, B., 2011. Extraction of narrative recall patterns for neuropsychological assessment. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech), pp. 3021–3024.
Rosenstein, M., Diaz-Asper, C., Foltz, P.W., Elvevåg, B., 2014. A computational language approach to modeling prose recall in schizophrenia. Cortex 55, 148–166. https://doi.org/10.1016/j.cortex.2014.01.021.
Ruff, R.M., 2003. A friendly critique of neuropsychology: Facing the challenges of our future. Arch. Clin. Neuropsychol. 18 (8), 847–864.
Schnabel, R., 2012. Overcoming the challenge of re-assessing verbal memory. Clin. Neuropsychol. 26 (1), 102–115.
Tal, A., Torous, J., 2017. The digital mental health revolution: opportunities and risks. Psychiatr. Rehabil. J. 40 (3), 263–265. https://doi.org/10.1037/prj0000285.
Topol, E., 2019. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st ed. Basic Books, Inc., USA.
Torous, J., Baker, J.T., 2016. Why psychiatry needs data science and data science needs psychiatry: connecting with technology. JAMA Psychiatry 73 (1), 3–4. https://doi.org/10.1001/jamapsychiatry.2015.2622.
Turney, P.D., Pantel, P., 2010. From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. 37, 141–188. https://doi.org/10.1613/jair.2934.
Wechsler, D., 1945. A standardized memory scale for clinical use. J. Psychol. 19, 87–95.
Wechsler, D., 1987. Wechsler Memory Scale - Revised. The Psychological Corporation, San Antonio, TX.
Wechsler, D., 1997. Wechsler Memory Scale - Third Edition, WMS-III: Administration and Scoring Manual. The Psychological Corporation, San Antonio, TX.
Wechsler, D., 2009. Wechsler Memory Scale - Fourth Edition, WMS-IV: Technical and Interpretive Manual. Pearson, San Antonio, TX.
Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2020. BERT score: evaluating text generation with BERT. In: International Conference on Learning Representations.