TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Sport- und Gesundheitswissenschaften

Lehrstuhl für Sportpsychologie

Construction and evaluation of a verbal memory test according to neurolinguistic criteria: the Auditory Wordlist Learning Test (AWLT)

Johannes Baltasar Heßler

Vollständiger Abdruck der von der Fakultät für Sport- und Gesundheitswissenschaften der Technische Universität München zur Erlangung des akademischen Grades eines Doktors der Philosophie (Dr. phil.) genehmigten Dissertation.

Vorsitzende:	Prof. Dr. Renate M. Oberhoffer
Prüfer der Dissertation:	Prof. Dr. Jürgen Beckmann
	apl. Prof. Dr. Thomas Jahn

Die Dissertation wurde am 08.01.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Sport- und Gesundheitswissenschaften am 19.07.2018 angenommen.

Acknowledgments

This dissertation carries my name, yet, is the result of many people's contributions. I thank Prof. Jürgen Beckmann for his ongoing support from the day I enrolled as a Ph.D. student and for welcoming me into his teaching staff.

My sincerest gratitude belongs to Prof. Thomas Jahn for allowing me to manage our project and granting me the freedom to do so in my own way. Also, I thank him for continuously fostering, guiding, and supervising me as a researcher. The tireless and meticulous effort he puts into his projects is an inspiration and I look forward to continuing our work on CFD and AWLT. I express my deepest appreciation to Dr. Horst Bickel for developing my scientific mind, unreservedly promoting my career, and confronting me with his view of psychiatry and academia.

I sincerely thank Mag. Gerald Schuhfried jun., Mag. Marco Vetter, Dr. David Brieber, Mag. Johanna Egle and Mag. Georg Mandler of the Schuhfried GmbH for giving me the opportunity to play my role in the CFD-project. Further, I am thankful to the staff of the Department of Psychiatry and Psychotherapy at the Klinikum rechts der Isar involved in our project, as well as all cooperating persons and institutions.

Most importantly, I am deeply grateful to those who are close to me for the support and love they have given me throughout my life.

Table of contents

I. Summary I

II. List of figures and tables III

III. Publication and submission record IV

1. Introduction 1

1.1 Structure of the thesis... 1

1.2 Dementia, mild cognitive impairment, and neurocognitive disorder................ 2

1.3 The role of neuropsychology in diagnosing dementia 4

1.4 The wordlist learning paradigm... 6

1.5 The test-set Cognitive Functions Dementia (CFD)... 7

1.6 The Auditory Wordlist Learning Test (AWLT) 7

2. Theoretical background 9

2.1 Verbal memory 9

2.2 Verbal memory and dementia 10

2.3 Linguistic recall effects in verbal memory 11

2.3.1 Word length 12

2.3.2 Word Frequency 13

2.3.3 Orthographic neighborhood size 14

2.4 Implications for wordlist learning tests 16

2.4.1 Interpretation of test scores.. 16

2.4.1 Construction of new tests... 17

2.5 Research objectives 18

3. Methods 19

3.1 Linguistic analysis with the dlexDB 19

3.2 Examining linguistic recall effects in a non-experimental setting.................. 20

3.3 Study designs and measures 23

3.3.1 Chapter 4 23

3.3.2 Chapter 5 24

3.3.3 Chapter 6 25

4. Differential linguistic recall effects in the California Verbal Learning Test in healthy aging and Alzheimer's dementia: analysis of routine diagnostic data 28

5. Applying psycholinguistic evidence to the construction of a new test of verbal memory in late-life cognitive decline: the Auditory Wordlist Learning Test 31

6. Linguistic fairness and differential validity of the Auditory Wordlist Learning Test (AWLT) in dementia of the Alzheimer's type 34

7. Discussion 37

7.1 Summary 37

7.2 Linguistic recall effects in clinical practice 38

7.3 Theoretical and clinical implications 39

7.4 Limitations and future research 40

7.5 Conclusion 41

Literature 43

Appendix 53

I. Summary

Abstract

Experimental evidence suggests that certain linguistic characteristics influence the probability by which a word is learned and recalled in wordlist learning tests and that these effects can present differently in persons with and without cognitive impairment. Analyzing routine data of memory clinic patients with and without dementia of the Alzheimer’s type (DAT), we found that these effects and interactions occurred in the German California Verbal Learning Test (CVLT). The German Auditory Wordlist Learning Test (AWLT) was, therefore, constructed based on linguistic criteria in order to mitigate linguistic recall effects and interactions, control the test’s linguistic item difficulty, and increase similarity within and between parallel forms. In a pilot study, the AWLT demonstrated good internal consistency and test-retest reliability. Using data from older persons without cognitive impairment and persons with DAT, we demonstrated a reduced susceptibility of the AWLT to linguistic interference. Further, the AWLT’s main variables showed high differential validity to discriminate between the diagnostic groups.

Zusammenfassung

Experimentelle Studien fanden, dass bestimmte linguistische Eigenschaften assoziiert sind mit der Wahrscheinlichkeit, mit der Wörter in Wortlisten Lerntests gelernt und erinnert werden und dass sich diese Effekte unterschiedlich darstellen für Personen mit und ohne kognitive Beeinträchtigung. Anhand von Routinedaten von Patienten mit und ohne Demenz vom Alzheimer-Typ (DAT) einer Memory Clinic konnten wir zeigen, dass ähnliche Effekte und Interaktionen im deutschen California Verbal Learning Test (CVLT) auftreten. Der Auditive Wortlisten Lerntest (AWLT) wurde entlang neurolinguistischer Kriterien entwickelt, um diese linguistischen Gedächtniseffekte zu verringern, die linguistische Itemschwierigkeit zu kontrollieren sowie die Ähnlichkeit innerhalb und zwischen den Wortlisten der Parallelformen zu erhöhen. In einer Pilotstudie bewies der AWLT gute interne Konsistenz und Test-Retest Reliabilität. Basierend auf Daten von älteren kognitiv gesunden Personen sowie Personen mit DAT zeigten wir eine verringerte Anfälligkeit des AWLT für linguistische Interferenz. Weiter unterschieden die Hauptvariablen des AWLT mit großen Effektstärken zwischen den diagnostischen Gruppen.

II. List of figures and tables

- FIGURE 1: Structure of the thesis. Page 1.
- FIGURE 2: Verbal memory processes according to Shacter and Wagner (2013). Page 9.
- TABLE 1: Overview of common linguistic terms used in the present thesis. Page 19.

III. Publication and submission record

This dissertation was submitted as a cumulative thesis and is based on the following three papers that have been published in or submitted to international peer-reviewed journals. The journals' permissions to include the already published papers in the dissertation can be found in the sections B and C of the appendix, respectively.

- Hessler, J., B., Fischer, A. M., & Jahn T. (2016). Differential linguistic recall effects in the California Verbal Learning Test in healthy aging and Alzheimer's dementia: analysis of routine diagnostic data. *Archives of Clinical Neuropsychology*, *31*, 689–699.

- Hessler, J. B., Brieber, D., Egle, J., Mandler, G., & Jahn, T. (2017). Applying psycholinguistic evidence to the construction of a new test of verbal memory in late-life cognitive decline: the Auditory Wordlist Learning Test. *Assessment*, online first.

- Hessler, J. B., Brieber, D., Egle, J., Mandler, G., & Jahn, T. (in print). Linguistic fairness and differential validity of the Auditory Wordlist Learning Test (AWLT) in dementia of the Alzheimer's type (article in German). *Diagnostica.*

1. Introduction

1.1 Structure of the thesis

The present thesis includes five main parts with seven chapters. *Figure 1* displays the overall structure of the thesis and details of the individual chapters.

FIGURE 1. Structure of the thesis.

1. Introduction	**2. Theoretical background**
• Structure of the thesis • Dementia, MCI, and NCD • The role of neuropsychology in the diagnosis of dementia • Cognitive Functions Dementia (CFD)	• Verbal memory • Linguistic recall effects • Implications for wordlist learning tests • Research objectives

3. Methods

- Linguistic analysis with the dlexDB
- Examining linguistic recall effects in a non-experimental setting
- Study designs and measures of chapters 4 – 6

Main studies

4. Differential linguistic recall effects in the California Verbal Learning Test in healthy aging and Alzheimer's dementia: analysis of routine diagnostic data
5. Applying psycholinguistic evidence to the construction of a new test of verbal memory in late-life cognitive decline: the Auditory Wordlist Learning Test (AWLT)
6. Linguistic fairness and differential validity of the Auditory Wordlist Learning Test (AWLT) in dementia of the Alzheimer's type

7. Discussion

- Summary
- Linguistic recall effects in clinical practice
- Theoretical and clinical implications
- Limitations and future research
- Conclusion

Literature	**Appendix**

Note. MCI = mild cognitive impairment, NCD = neurocognitive disorder

Introduction (1) and theoretical background (2) outline the research context, in which the dissertation is situated, as well as its objectives. The methods (3) of the three main studies are subsequently described and it is reflected on how linguistic recall effects can be examined in a non-experimental setting with the linguistic database dlexDB. The three studies are then presented with their abstracts (4 – 6). The overall findings are summarized in the discussion (7) and their theoretical as well as clinical implications are outlined, along with overall limitations of the dissertation, and an outlook on future research is given. The dissertation is concluded with the references and the appendix, which comprises the published or submitted full texts of the main studies.

1.2 Dementia, mild cognitive impairment, and neurocognitive disorder

Dementia describes the deterioration of any cognitive function compared to an earlier point in time to the degree that daily functioning is impaired. Traditionally, memory impairment was emphasized in diagnostic criteria, as it represents the cardinal symptom of dementia of the Alzheimer's type (DAT), the most common form (Reitz, Brayne, & Mayeux, 2011). Next to Alzheimer's disease, which is considered the main cause for DAT, a wide range of underlying etiologies for dementia have been identified, for example, vascular impairment (Iadecola, 2013) and frontotemporal lobar degeneration (Riedl, Mackenzie, Förstl, Kurz, & Diehl-Schmid, 2014). While etiology plays only a secondary role in diagnosing dementia, as it must not necessarily be stated, its correct identification is paramount for weighing treatment options and prognosis. The incidence of dementia before the age of 65 is relatively low (Harvey, Skelton-Robinson, & Rossor, 2003) but the number of new cases increases exponentially in the years after (Jorm & Jolley, 1998). In 2013 (Prince et al.), it was estimated that worldwide 35.6 million people live with dementia and that this number will double every 20 years.

Attempts to define the grey area between healthy aging (i.e., cognitive impairment that would be expected for a certain age) and dementia, which is signified by reduced cognitive functioning while daily activities are retained, resulted in a variety of constructs, of which mild cognitive impairment (MCI; Gauthier et al., 2006; Winblad et al., 2004) was most endorsed in the recent years. The consensus criteria proposed by Winblad and colleagues (2004) include (1) a state of being not normal, not demented (does not meet criteria (DSM-IV, ICD 10) for a dementia syndrome); (2) cognitive decline specified by self and/or informant report and impairment on objective cognitive tasks and/or evidence of decline over time on objective cognitive tasks; and (3) preserved basic activities of daily living or minimal impairment in complex instrumental functions.

While also suffering from conceptual problems (e.g., Hessler, Tucha, Förstl, Mösch, & Bickel, 2014b), MCI actually describes a very heterogeneous group with a wide range of etiologies and, as a result, differential disease trajectories (Mitchell & Shiri-Feshki, 2008; 2009). As some persons diagnosed with MCI may progress to dementia, the concept rather possesses clinical utility than diagnostic validity (Gainotti, 2010).

Diagnosing dementia according to diagnostic manuals proceeds in two steps. First, the general symptoms of a dementia syndrome need to be recognized in a patient. Second, the probable underlying cause is determined. The ICD-10 (World Health Organization, 1993) offers no diagnostic code for dementia in general and affords the clinician to directly code the diagnosis according to the most likely etiology. DSM-IV (American Psychiatric Association, 2000) and DSM-5 (American Psychiatric Association, 2013), in turn, allow for coding a general dementia syndrome and, in a second step, the underlying etiology. The DSM-5 introduced the term Neurocognitive Disorder (NCD), whose mild and major forms conceptually correspond to MCI and dementia, respectively. The definitions according to the three diagnostic manuals are shown in *Table a* in the appendix. For the remainder of the dissertation, the terms dementia and MCI will by favored to NCD, as the former have established themselves in the recent years and were employed by most research cited here.

1.3 The role of neuropsychology in diagnosing dementia

Mirroring the approach of the manuals, the diagnostics of dementia follow a two-step process (Jahn & Werheid, 2015). First, the presence of a dementia syndrome needs to be established and, second, the underlying etiology is specified. Neuropsychological tests are especially important in the first step, when the level of present and past cognitive functioning needs to be determined, however, they also add valuable information in the second step.

The assessment of cognitive functioning is central in the diagnostics of dementia, as there needs to be evidence of deterioration compared to a previous point in time. Neuropsychological tests are primarily employed to assess current cognitive functioning. Premorbid functioning is most accurately estimated from variables that remain unaffected by dementia, like age, gender, and education (Jahn et al., 2013). Direct neuropsychological measures of premorbid functioning, for example tests of verbal intelligence, have been proposed to be robust against mild and moderate cognitive impairment, however, the opposite was repeatedly demonstrated (Binkau, Berwig, Jänichen, & Gertz, 2014; Hessler, Jahn, Kurz, & Bickel, 2013). The main role of neuropsychological testing is, therefore, to quantify current deficits in a range of cognitive domains and determine whether these deficits are sufficiently pronounced to be considered clinically relevant. Neuropsychological diagnostics primarily work on a symptomatic level. Certain patterns of deficits, however, can discriminate between different etiologies of dementia (Beck, Schmid, Berres, & Monsch, 2013; Jahn & Werheid, 2015).

Depending on the setting, cognitive tests of varying length and psychometric quality are employed. Screening tests such as Mini Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975), Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005), or the 6-Item Cognitive Impairment Test (6CIT; Brooke & Bullock, 1999; Katzman et al., 1983) are short measures of global cognitive functioning that are useful when there is little time and a rough estimate of cognitive status is needed (Hessler et al., 2016b; 2014a). Due to their briefness and lack of differentiation, these screenings are by no means suited to justify a diagnosis of dementia.

Two general approaches to neuropsychological assessment also find application in the context of dementia. Standardized batteries present preselected tests in a fixed order that are usually co-normed and allow for time-efficient testing. The most common specimen is the neuropsychological test battery of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD-NTB; Morris et al., 1989). Inherent to this approach, however, comes an inflexibility with regard to specific diagnostic questions. As an alternative, the flexible battery approach (FBA) describes the compilation of individual tests into a case-specific test-battery. While this approach allows for a high adjustment to specific diagnostic questions and patient variables, problems with scoring and interpretation may arise from different norms and their quality. Standardized test batteries like the CERAD-NTB are usually favored, as they balance a sufficient amount of diagnostic information with relatively short administration times. Further, the FBA affords a deeper understanding of neurocognitive testing and the neuropsychology of dementia, so that it might not be feasible in less specialized settings.

While DSM-IV and ICD-10 do not include such recommendations, the DSM-5 for the first time demanded standardized neuropsychological testing, described explicit neurocognitive domains of possible impairment, alluded to specific neuropsychological tests, and suggested that diagnostic decisions should be based on quantitative test variables. Importantly, neuropsychological tests results always need to be considered in the context of physiological measures including analyses of blood and cerebrospinal fluid, as well as brain imaging techniques, evidence of daily functioning, informant reports, and the exclusion of other conditions like delirium or depression in order for a diagnosis of MCI or dementia to be made (Jahn & Werheid, 2015).

1.4 The wordlist learning paradigm

Tests of (verbal) memory are a cardinal part of every neuropsychological dementia assessment. The wordlist learning paradigm is commonly employed in this context. The paradigm is characterized by a certain succession of subtests that aim to assess different aspects of verbal memory. Initially, during the learning phase, the patient is presented a list of words in some form. The words can be read aloud by the clinician or shown to the test person. Some tests require the patient to read the list out to ensure that the words are encoded. Immediately after presentation, the words are to be recalled, usually in an unstructured way. This sequence of presentation and immediate recall is repeatedly conducted, usually between three and five times. During the delayed recall phase, the words are to be named again after a short and/or a long delay, which may be conducted with or without giving semantic or lexical cues to promote recall. Usually, the patient is not previously informed that a delayed recall will be asked for. The subsequent recognition subtest requires the test person to identify the learned words among distractors.

1.5 The test-set Cognitive Functions Dementia (CFD)

The revaluation of neuropsychological assessment and the specification of potentially impaired cognitive domains in the DSM-5 indicated a re-evaluation of the testing landscape. The predominant CERAD-NTB yields only little information on attention and executive functioning, includes subtests of questionable psychometric quality (Katsumata et al., 2015), and gives away diagnostic information through limited scoring possibilities. The FBA is adaptable to these changes, however, might not be feasible in certain settings. In the light of this situation, the Schuhfried GmbH (Mödling, Austria) and the Clinical and Experimental Neuropsychology Unit of the Department of Psychiatry and Psychotherapy at the Klinikum rechts der Isar of the Technical University of Munich (Prof. Thomas Jahn) cooperated in the construction of the test-set Cognitive Functions Dementia (CFD; Jahn & Hessler, 2017) that is run on the Vienna Test System (Wiener Testsystem, WTS). The CFD assesses the cognitive domains specified by the DSM-5 (except social cognition); is fully administrated on a device with touchscreen; can be employed in a standard-, long, or screening-version; exists in two parallel forms for repeated testing; is normed for persons aged 50 years or older; and includes computerized test scoring and individual case statistics for repeated measurements. The CFD was released in 2017 and is currently being validated in a multicenter study on patients with different forms of MCI and dementia.

1.6 The Auditory Wordlist Learning Test (AWLT)

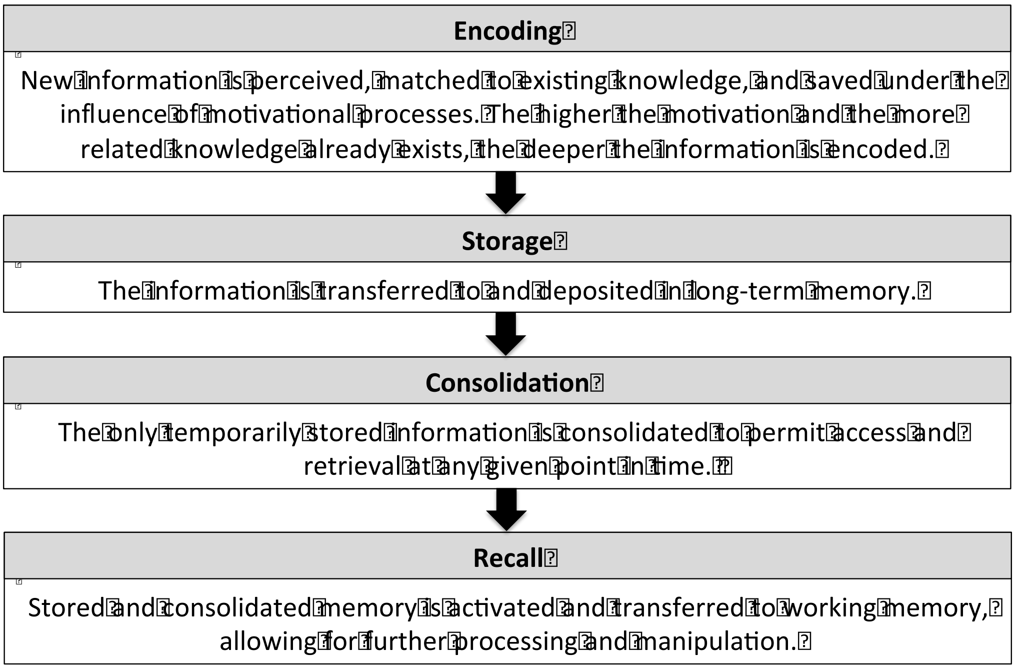
The Auditory Wordlist Learning Test (AWLT; Hessler & Jahn, 2017) was constructed as a central part of the CFD in order to assess verbal memory. The AWLT follows the described wordlist learning paradigm and was constructed according to a novel neurolinguistic approach. The present dissertation describes the clinical and theoretical basis for this approach, the AWLT’s construction, and its evaluation with regard to linguistic fairness as well as differential validity.

2. Theoretical background

2.1 Verbal memory

Verbal memory is a subdivision of episodic memory, which contains recollections of specific situations from a person's life that are stored together with information about the time and date they took place at (Shacter & Wagner, 2013). Accessing their episodic memory, most people, for example, can recall where they were and what they did when the World Trade Center collapsed in 2001. Together, episodic memory and semantic memory, the memory for facts that are stored without information about the situation they were acquired in, constitute the explicit long-term memory system. Four processes describe how explicit long-term memories, including verbal memories, are formed and recalled (*Figure 2*). Disruption of any of these processes may lead to memory impairments.

FIGURE 2. Verbal memory processes according to Shacter and Wagner (2013).



Encoding

New information is perceived, matched to existing knowledge, and saved under the influence of motivational processes. The higher the motivation and the more related knowledge already exists, the deeper the information is encoded.

Storage

The information is transferred to and deposited in long-term memory.

Consolidation

The only temporarily stored information is consolidated to permit access and retrieval at any given point in time.

Recall

Stored and consolidated memory is activated and transferred to working memory, allowing for further processing and manipulation.

These processes are assumed to have specific neural correlates (Shacter & Wagner, 2013). After auditory or visual sensory processing, the verbal information is transmitted to the phonological loop, a subunit of the working memory. Here, rehearsal regulated by Broca's area keeps the information alive, while it is being matched to existing semantic knowledge through activation of the posterior parietal lobe. The information is then transferred to long-term memory, which, together with subsequent recall and recognition processes, is mediated by medial temporal lobe, hippocampus, and left prefrontal cortex.

2.2 Verbal memory and dementia

Memory deficits represent the hallmark symptom of DAT (Bondi et al., 2008) and may also occur in other forms, especially vascular dementia (Erkinjuntti, Laaksonen, Sulkava, Syrjäläinen, & Palo, 1986). DAT and vascular dementia represent the two most common forms, together accounting for around 75% of dementia cases (Fratiglioni, De Ronchi, & Agüero-Torres, 1999).

Of the memory domains, episodic memory functioning, and with that verbal memory, seems to be most affected in DAT. This finding is readily explained by the typical progression of the Alzheimer neuropathology affecting already at early stages the medial temporal lobe, including entorhinal and transentorhinal areas as well as the hippocampus (Bondi et al., 2008; Wolk & Dickerson, 2011). Accordingly, wordlist learning tests show some of the largest effect sizes for differentiating between healthy persons and persons with DAT (Han, Nguyen, Stricker, & Nation, 2017; Zakzanis, Leach, & Kaplan, 1999). In other dementias that are not characterized by early memory impairment, such as frontotemporal dementia and dementia due to Parkinson's disease, verbal memory is usually retained for longer periods. However, such impairment may occur at later stages when neuropathologic changes affect multiple brain areas and cognitive deficits generalize (Jenner & Benke, 2002).

Testing for memory impairment is an essential part of every neuropsychological dementia assessment. As memory deficits already occur in prodromal phases of DAT (Bäckman, Jones, Berger, Laukka, & Small, 2005) and have high predictive validity for the progression from MCI to dementia (Gainotti, Quaranta, Vita, & Marra, 2014), their detection is instrumental for the early identification of DAT. Further, early dementia needs to be differentiated from geriatric depression, which may be accompanied by cognitive deficits. While these symptoms tend to remit after successful antidepressive treatment (Butters et al., 2000), the cognitive symptoms of dementia can only be delayed in their progression by psychosocial interventions (e.g., Vernooij-Dassen, Vasse, Zuidema, Cohen-Mansfield, & Moyle, 2010) and pharmacological treatment (e.g., Raina et al., 2008). Measures of verbal memory have been repeatedly shown to successfully discriminate between persons with dementia and depression (e.g., (Foldi, Brickman, Schaefer, & Knutelska, 2003; Jahn et al., 2004).

2.3 Linguistic recall effects in verbal memory

Verbal memory functioning in the wordlist learning paradigm partly depends on the characteristics of the words to be learned (Rubin & Friendly, 1986). These effects have been demonstrated in several languages (Brysbaert et al., 2011) and, therefore, seem to reflect universal processes of verbal memory. For the purpose of the construction of the AWLT, word length, frequency, and neighborhood size were considered in more detail, even though there are more potentially relevant factors, such as age of acquisition (the age at which a word is first encountered and learned; Zevin & Seidenberg, 2002), familiarity (a vague feeling of having encountered a word before; Yonelinas, 2002), or imaginability (the degree to which a word can be mentally depicted; Paivio, Yuille, & Madigan, 1968). We decided to take into account only the three named characteristics, as they appeared to be most relevant based on the existing literature, while the others might be more important in tasks of picture naming, reaction time, and speech production. Further, considering too many linguistic factors in the word selection for the AWLT would reduce the number of eligible words and likely require to deviate from the AWLT's construction criteria in order to fill the learning lists with words. Consequently, word length, frequency, and neighborhood size were chosen as highly relevant and also easily quantifiable linguistic characteristics.

2.3.1 Word length

Orthographic word length (as opposed to phonological word lengths, which depends on the number of phonemes or the time it takes to pronounce a word) can be operationalized as the number of syllables or letters of a word (Jalbert, Neath, Bireta, & Surprenant, 2011b). For the construction of the AWLT, orthographic word length was chosen as easily determinable measure and operationalized as the number of syllables.

The so-called word length effect is based on the often replicated finding that, when comparing recall rates for pure lists containing only short words with recall rates for pure lists of long words, more short words are recalled (Jalbert, Neath, Bireta, & Surprenant, 2011b). In mixed lists with both short and long words, however, there seems to be no difference in recall rates (Bireta, Neath, & Surprenant, 2006; Jalbert, Neath, Bireta, & Surprenant, 2011b). These findings suggest that working memory has limited phonological capacity, which can hold multiple short but only few long words. This notion contributed to the integration of the phonological loop in working memory models (Baddeley, Thomson, &

Buchanan, 1975) and leads to the prediction that learning list with mainly short words would produce higher recall scores that lists with mainly long words.

2.3.2 Word Frequency

Word frequency refers to the commonness of a word in its language. Called the word frequency mirror effect or word frequency paradox, a common finding in wordlist learning studies is that cognitively healthy persons tend to learn and recall frequent words better, while infrequent words are better recognized among distractor words (Criss, Aue, & Smith, 2011; DeLosh & McDaniel, 1996; MacLeod & Kampe, 1996). Importantly, this effect was only observed when comparing memory performance with pure lists of either frequent or infrequent words. Studies employing mixed lists that include both frequent and infrequent words also lead to mixed results, suggesting an advantage for frequent (e.g., Criss et al., 2011) but also for infrequent words (e.g., Ozubko & Joordens, 2007). This equivocality may be explained by parametric studies that examined the association between frequency and recall probability in a large number of words. The results indicated a U-shaped association with both frequent and infrequent words having high recall rates, possibly explained by the notion that both groups are particularly salient through their non-normal frequency (Lohnas & Kahana, 2013).

Presumably, the word frequency mirror effect can be explained by a process called redintegration (Roodenrys, Hulme, Lethbridge, Hinton, & Nimmo, 2002), which describes the reconstruction of the whole from a fragment. In psycholinguistics, the whole is represented by the word to be recalled and the fragment by its decaying memory trace. Frequent words are assumed to leave a stronger memory trace that is better picked up during recall than the trace of infrequent words. The preference for frequent words is also reflected in other basal neurocognitive levels. For example, frequent words are faster processed and, therefore, have shorter fixation times during reading (Inhoff & Rayner, 1986) and afford less energy during encoding (Diana & Reder, 2006).

Clinical studies suggested possible interactions between word frequency and the cognitive status of the participants. In person sedated with alcohol and lorazepam the advantage for frequent words was significantly diminished compared to a placebo control group (Soo-ampon, Wongwitdecha, Plasen, Hindmarch, & Boyle, 2004). In a study comparing the recall of frequent and infrequent words in persons with schizophrenia and healthy controls, however, no difference was found (Brébion, David, Bressan, & Pilowsky, 2005).

In a word recognition task that affords to identify previously learned words among distractors, persons with DAT showed a reduced hit rate (i.e., correct identification of target words) for infrequent words compared to healthy controls, while both groups had similar hit rates for frequent words (Balota, Burgess, Cortese, & Adams, 2002; Wilson, Bacon, Fox, Kramer, & Kaszniak, 2008). A possible explanation might be that the two groups employ different memory processes in order to decide which word to rule in or out. While controls might use recollection (i.e., explicit recognition of a specific word) to identify the salient infrequent words, persons with DAT potentially use the vague feeling of familiarity to identify the frequent words (Yonelinas, 2002). In sum, the results suggest that persons with cognitive impairment may not benefit from the same linguistic recall effects as healthy persons and, therefore, need to employ other strategies in the effort to promote memory performance.

2.3.3 Orthographic neighborhood size

Orthographic neighborhood size according to Levenshtein's algorithm (Levenshtein, 1966) refers to the number of meaningful words that can be derived from the addition, removal, or replacement of a single letter in a word. Importantly, for every neighbor, only one operation may be performed. For example, the words “bloat”, “boa”, and “goat” are all neighbors of “boat”. Experimental studies showed that words with larger neighborhood sizes (i.e., more neighbors) were better recalled than words with small neighborhood sizes (Allen & Hulme, 2006; Jalbert, Neath, & Surprenant, 2011a; Jalbert, Neath, Bireta, & Surprenant, 2011b; Roodenrys et al., 2002). Similar to word frequency, this effect might be explained by redintegration, suggesting that words with more neighbors leave stronger memory traces and, thereby, promote recall (Jalbert, Neath, Bireta, & Surprenant, 2011b).

An effect similar to the word frequency paradox became apparent in recognition tasks. Words with fewer neighbors are usually better recognized than words with more neighbors (Glanc & Greene, 2007). Words with smaller neighborhood sizes appear to be linguistically more salient and, therefore, are easier recognized as they can be consciously recollected (as opposed to recognition judgments based on familiarity). These findings suggest that words with different neighborhood sizes rely on different cognitive processes in verbal memory tasks. So far, there is no evidence on recall effects of neighborhood size in clinical settings.

It has been argued that the effect of neighborhood size on learning and memory actually underlies the word length effect (Jalbert, Neath, & Surprenant, 2011a; Jalbert, Neath, Bireta, & Surprenant, 2011b). As short words have more neighbors it is possible that neighborhood size mediates the association between word length and recall rates. When equating words of different length with regard to their neighborhood sizes the word length effect vanished in two studies (Jalbert, Neath, & Surprenant, 2011a; Jalbert, Neath, Bireta, & Surprenant, 2011b).

2.4 Implications for wordlist learning tests

So far the evidence on linguistic recall effects in clinical populations is scarce. To our knowledge, no studies have examined linguistic recall effects in clinically employed wordlist learning tests. This is surprising, given that the experimental evidence has several implications for existing tests. Further, the linguistic knowledge may guide the construction of new tests aiming to reduce or even prevent the limitations caused by linguistic interference.

2.4.1 Interpretation of test scores

Depending on the linguistic composition of a wordlist, interactions between linguistic characteristics and the test-taker's cognitive status may become diagnostically relevant. If, for example, the wordlist included mostly frequent words, a disadvantage for persons with cognitive impairment might arise, since they might not be able to capitalize on the word frequency effect as cognitively healthy persons usually do. As a result, group differences might be larger than in a linguistically "fair" list and the severity of impairment might be overestimated. A list containing mostly words with small neighborhood sizes might have a reduced ability to discriminate between persons without and with DAT. While the former might anyways recall more words, the latter potentially benefit from the word's linguistic salience and show better recall scores. In lists with mostly short words, ceiling-effects are more likely, as more items can be kept active in working memory between learning and recall.

In addition, linguistic recall effects potentially mask or imitate diagnostically important position effects during immediate and delayed recall. Cognitively healthy persons usually show primacy- and recency effects in recall with more words being recalled from beginning and end of the list. These position effects are important parameters for differential diagnostics in dementia. While persons with depression show both primacy- and recency-effects, persons with dementia often display a reduced primacy-effect with a retained or even more pronounced recency-effect (Harris & Dowson, 1982). In addition, a reduced primacy-effect in persons with MCI was associated with an increased risk of progressing to dementia (Cunha, Guerreiro, de Mendonça, Oliveira, & Santana, 2012; Egli et al., 2014). Since frequency effects seem to be even stronger at recency-positions (Van Overschelde, 2002), an accumulation of frequent words at the end of the learning list may lead to a reduced primacy- and increased recency-effect even in healthy persons. As a consequence, the diagnostic validity of parameters based on serial position effects would be compromised.

The exact knowledge of how linguistic effects interact with different presentations of cognitive impairment would allow for the calculation of psycholinguistic test-parameters. For example, based on experimental evidence, the difference in recall rates for high frequency minus low-frequency words might be a sensitive marker for cognitive changes in DAT. Persons with dementia would be expected to have scores that tend to approach 0. The evidence-base, however, is by far not large and unambiguous enough to exhaustively understand and reliably predict these interactions. Therefore, the more favorable approach is to reduce these interactions as much as possible and create a linguistically fair wordlist learning test.

2.4.1 Construction of new tests

The construction of new wordlist learning tests should be informed by the evidence and reasoning presented above. That is, words should be selected according to linguistic criteria that aim to prevent floor- and ceiling-effects, avoid to mask or imitate position effects, and create a linguistically fair wordlist that does not produce interactions between linguistic composition and cognitive status. Further, the linguistic knowledge could be used to influence the item difficulty of a wordlist learning test and to create parallel forms that are similar even at the item level.

2.5 Research objectives

The overall aim of the dissertation project was threefold: (1) To investigate whether interactions between linguistic characteristics and the test takers' cognitive status are present in wordlist learning tests that are used in clinical practice. (2) To construct a new wordlist learning test according to linguistic criteria based on the evidence from (1) and the available literature. (3) To examine whether the newly constructed test is actually linguistically fair and able to assess a range of aspects of verbal memory in persons with and without dementia as well as differentiate between these groups. The studies conducted to reach these aims are described in the chapters 4, 5, and 6, respectively.

3. Methods

3.1 Linguistic analysis with the dlexDB

All linguistic analyses in the present thesis were performed with the dlexDB (Heister et al., 2011). *Table 1* gives an overview of the most important linguistic terms related to the linguistic analyses in the present thesis.

TABLE 1. Overview of common linguistic terms used in the present thesis.

Token
Every occurrence of a word in a lexical database is represented by a token. If a certain word occurs twice, it generates two tokens.

Type
Every first occurrence of a certain word. If the same word occurs several times, it generates one type.

Annotation
Additional information for a type, indicating the part of speech it belongs to.

Normalization
Method of standardization, issuing the count of a word in the corpus per one million tokens or types in the corpus.

Type frequency
The frequency of a specific type in the corpus.

Orthographic neighborhood size (Levenshtein, 1966)
The number of words that can be derived by adding, removing, or replacing a single letter in a certain word, while per neighbor only one operation can be performed.

Normalized annotated type frequency
The standardized type frequency differentiated according to their part of speech.

Normalized neighborhood size
The standardized number of orthographic neighbors in the corpus.

Orthographic length
The number of a word’s syllables.

The dlexDB is a linguistic database that builds on the text corpus of the Digitales Wörterbuch der deutschen Sprache (DWDS; Geyken, 2007), which includes 122.816.010 tokens (words in total including repetitions) and 2.224.542 types (individual words) and, thereby, constitutes the largest German text corpus. Its words were extracted from fiction, newspaper articles, functional texts, prose, and transcribed spoken language from the whole 20th century. The dlexDB is accessible online (www.dlexdb.de), free of charge, and offers a filter to select variables for the detailed linguistic analysis of single words and word lists.

Analyses with the dlexDB were performed in order to examine the German California Verbal Learning Test (CVLT; Niemann, Sturm, Thöne-Otto, & Willmes, 2008) with regard to its linguistic profile (chapter 4) and to select the AWLT's words (chapter 5).

3.2 Examining linguistic recall effects in a non-experimental setting

As described in the introduction, a range of experimental studies examined the effect of a word's characteristics on its probability of being learned, recalled, and recognized. Inherent to the experimental setting is a need for strong control over the stimuli (i.e., the words) presented to the participants. With databases like the dlexDB, words can be matched with regard to a range of their linguistic characteristics to isolate the effects of other characteristics that can then be systematically varied. This approach is not feasible in a clinical setting. Given that they are employed as diagnostic instruments rather than experimental stimuli, wordlists developed for clinical purposes are usually not balanced or strictly controlled with regard to their linguistic characteristics. Further, when examining existing lists, there is no control over the stimuli. As a consequence, the experimental approach to examining linguistic recall effects needs to be adapted to the applied clinical setting.

Neuropsychological examinations of patients involving wordlist learning tests could be considered within-subjects quasi-experiments with linguistic conditions that are defined by the make-up of the respective wordlist. The individual words of the learning list could be sorted into a certain number of experimental conditions, for example, high frequency versus low frequency words. An obvious way to do so would be to consult established criteria that discriminate between frequent and infrequent words. Since there are no such criteria that define frequency in an absolute manner, a measure of relative frequency within that specific list needs to be applied. We chose the median as the cut-off to sort words into categories, as the mean might be too amenable to potential outliers. The median, in contrast, separates the word-list into two roughly equally sized word groups that can be assumed to differ in their respective mean frequencies. The groups' labels "frequent" or "infrequent" are consequently to be understood as "relative to the other group". As words at the beginning and the end of a list are usually better learned and recalled (Murdock, 1962), possible mediating or masking effects of the words' serial position need to be investigated. Assuming that words are not sorted according to their linguistic features within the list, no interaction between serial position and linguistic characteristic would usually be expected.

When completing the wordlist learning test, the patient is virtually exposed to a range of randomly sorted stimuli that belong to one of the two groups defined by the median. To examine the effect on learning and recall of the two experimental conditions, the percentage of recalled words in each condition is calculated. Group differences in mean recall rates can then be statistically compared in order to determine whether linguistic recall effects are relevant for this specific wordlist learning test.

The described method has high statistical power and flexibility, as it defines patients as cases and words as within-subjects variables in the sense of a condition or time variable. This set-up allows for employing repeated measures models that can be adjusted to the question at hand. Between-subjects variables can be introduced (e.g., diagnosis: patients versus controls), further within-subjects variables can be introduced (e.g., time: learning trial 1 versus short delay recall versus long delay recall), and covariates like age, gender, and education can be added. Examining the respective main and interaction effects can answer a range of research question, for example, is there a difference in recall rates between frequent and infrequent word and does this effect present differently between diagnoses and over time. The alternative approach would be to consider words as cases and simply correlate linguistic characteristics as quantitative variables with their respective recall rates in a given sample. This method, however, suffers from low power, as the number of cases is defined by the number of words in the list. The CVLT's learning list, for example, contains 16 words. Increasing the number of persons tested would only lead to better estimates of recall rates, but not to higher power. Also, more elaborate analyses are hardly possible, as they would entail dividing the small number of words even further.

Chapter 4 describes a study that was performed with routine diagnostic data of a memory clinic in order to investigate whether psycholinguistic findings from experimental studies also transfer to the applied setting and clinical instruments. With this design come a predefined set of stimuli (i.e., the wordlist of the employed test) and little control over the participants' characteristics. While these factors need to be accepted as inherent limitations to a "real-life" study, the described approach allows for structuring and analyzing the resulting data in a quasi-experimental manner. To our knowledge, this approach has been proposed for the first time in the study described in chapter 4.

3.3 Study designs and measures

3.3.1 Chapter 4

Chapter 4 (Hessler, Fischer, & Jahn, 2016a) describes how linguistic recall effects were for the first time investigated in a clinical sample with a diagnostic instrument.

Participants were retrospectively identified from archived patient records of the Department of Psychiatry and Psychotherapy at the Klinikum rechts der Isar of the Technical University of Munich. These patients were initially examined at the department’s Center for Cognitive Disorders (Zentrum für Kognitive Störungen) with suspected neurodegenerative disease and underwent cognitive testing at the department’s neuropsychological unit between January 1998 and August 2008. We selected patients that were discharged without a cognitive diagnosis (controls) or who were diagnosed with DAT for the first time. A subsample was created by matching controls with those patients with DAT based on age, gender, education, and depressive symptoms.

Diagnoses at the memory clinic were based on neuropsychological testing with the German CERAD-NTB (Tahlmann & Monsch, 1997), structural and/or functional neuroimaging, and cerebrospinal fluid diagnostics. At the neuropsychological unit, a FBA including the German CVLT was employed (Jahn et al., 2004). All statistical analyses were based on these routine data.

Linguistic analyses were performed as described above. Word length was operationalized as the number of a word’s syllables. Word frequency and neighborhood size were determined with the dlexDB. Annotated normalized type frequency was selected as a parameter indicating word frequency and normalized neighborhood according to Levenshtein’s algorithm (Levenshtein, 1966) as indicator for neighborhood size. For each linguistic characteristic, the

CVLT's learning list was dichotomized at the respective median. Percentages of recalled words for each linguistic condition (above or below median) were then calculated at learning trials 1 and 5 as well as long delayed free recall. For each linguistic characteristic, we performed a separate 2 (*diagnosis*: control vs. DAT) × 3 (*time*: trial 1 vs. trial 5 vs. long delay free recall) × 2 (*linguistic*: low vs. high) repeated measures analysis of variance (RM-ANOVA) of recall rates with *time* and *linguistic* as within-subjects factors and *diagnosis* as between-subject factor. Post-hoc *t*-tests were only calculated for significant interaction effects involving linguistic factors.

3.3.2 Chapter 5

Chapter 5 (Hessler, Brieber, Egle, Mandler, & Jahn, 2017) covers the AWLT's development, which proceeded in four steps. Development of a beta version, pilot testing of the beta version, test adjustment, and pilot testing of the final version.

The AWLT was developed along neurolinguistic criteria to control the test's difficulty on the item level, increase similarity between parallel forms, and reduce interactions between linguistic variables and the cognitive status of the test takers. The AWLT is based on the established wordlist learning paradigm, which has been previously described. Word selection for the learning list followed a five-step sequence. First, 12 semantic categories were defined to which the AWLT's words were to belong. Second, a pool of one- and two-syllabled words belonging to these categories was created. Third, these words were analyzed with regard to frequency and neighborhood size with the dlexDB. Fourth, 48 words were selected from the pool with normalized frequency and normalized neighborhood size between 5 and 15. Fifth, these words were distributed across four lists (learning and distraction for two parallel forms) in a way that each semantic category appeared only once on each list and that the average values for frequency and neighborhood size of each list lay around 10.

In a study with cognitively healthy subjects, who were registered in a database for study volunteers, in the Schuhfried testing center this first version of the AWLT was piloted in a paper and pencil format. Based on results of this study, changes were applied to the AWLT's learning list and the final version was created.

The final version was then again tested in a second pilot study with cognitively healthy participants in the Schuhfried testing center. With data from the second pilot study, reliability coefficients and mean recall differences between subtests were calculated. Further, similar to the study in chapter 5, linguistic recall effects were investigated.

3.3.3 Chapter 6

Chapter 6 (Hessler, Brieber, Egle, Mandler, & Jahn, under review) examined the presence of interactions between the test takers' cognitive status and the linguistic characteristics of the AWLT. Further, the validity of the AWLT's main variables to differentiate between older persons with and DAT was investigated.

Participants were aged 50 or older and tested with the CFD in the course of its norming and validation. The norming study was conducted in Munich (Technical University of Munich) and Vienna (Schuhfried GmbH). Participants were recruited in cooperation with retirement homes, municipal centers for older people, family centers, adult education centers (Volkshochschule), ads in newspapers and online forums, distribution of flyers on the street and in pharmacies, and personal contacts of the researchers. The participants volunteered by reaching out to the institutions and were then screened on the phone. Exclusion criteria were (1) age below 50 years; (2) insufficient proficiency of the German language; (3) currently in treatment of a neurological or psychiatric condition; (4) life-time incidence of stroke, severe head injury, severe traumatic brain injury, or meningitis; (5) currently in chemo- or radiation therapy; and (6) delirium/acute confusion within the last five years. Eligible persons were tested at the Klinikum rechts der Isar or the Schufried testing center. In some cases, participants were tested in their homes or at the centers they were recruited from.

The validation of the CFD is an ongoing study. The data employed in chapter 6 was collected in 9 clinical institutions. Here, the CFD was incorporated in the respective diagnostic processes and diagnoses of cognitive disorders were assigned according to each institution's protocols. At the time of writing chapter 6, the validation study only ran for several months. As a consequence, the clinical sample is relatively small and includes heterogeneous diagnoses that result from the institutions specialties. For the purpose of the present study, all patients with DAT were selected and matched to healthy controls from the CFD norming sample based on age, education, and gender.

The investigation of the AWLT's linguistic fairness basically followed the methods described for the related research questions in chapters 4 and 5. The AWLT's learning list was dichotomized three times at the median of the words' orthographic length, frequency, and orthographic neighborhood size, respectively. For each linguistic characteristic, a 2 × 3 × 2 RM-ANOVA was conducted with the between-subjects factor *diagnosis* (unimpaired vs. DAT) and the within-subjects factors *subtest* (learning trial 1 vs. learning trial 4 vs. long delayed recall) as well as recall rates according to *linguistic characteristic* (high vs. low).

The differential validity of the AWLT's main variables learning sum, short delayed recall, long delayed recall, and recognition as well as the Index Verbal

Long-term Memory from the CFD were examined by calculating Cohen's *d* effect sizes for group mean comparisons and performing receiver operator characteristics analyses for comparisons between healthy persons and persons with DAT.

4. Differential linguistic recall effects in the California Verbal Learning Test in healthy aging and Alzheimer's dementia: analysis of routine diagnostic data

Authors	Johannes Baltasar Hessler, Alina Maria Fischer, Thomas Jahn
Publications status	published in 2016 in *Archives of Clinical Neuropsychology*, *31*, 689–699.
Copyright	Oxford University Press
Location in appendix	B
Individual contribution	The Ph.D.-candidate is the main and first author of the paper, developed the research idea, chose the research design, analyzed the data, and conducted the submission and revision process, all under the supervision of and in agreement with Prof. Dr. Thomas Jahn.

Abstract

The aim of this study was to investigate the presence of linguistic recall effects in a clinical setting with a clinical instrument, the German CVLT. So far, these effects have only been demonstrated in experimental settings with high control over list compilation and participant selection. In clinical settings, the presence of linguistic recall effects, especially interactions with cognitive status of the test-takers, might compromise the test's diagnostic fairness and validity.

Employing routine diagnostic data extracted from archives of the Department of Psychiatry and Psychotherapy of the Klinikum rechts der Isar, Technical University of Munich, the effects of length, frequency, and neighborhood size of the CVLT's words on recall rates in controls and patients with DAT were investigated with repeated measures analysis of variance (RM-ANOVA) for the learning phase trials 1 and 5, as well as the long delayed free recall.

The results indicate that word length had no effect on recall rates in learning and recall phases of the CVLT. Word frequency had a main effect on recall rates aggregated across diagnosis and time of recall with high-frequency word having higher mean recall rates. Also, the interaction between frequency and time of recall reached statistical significance. Post-hoc *t*-tests suggested better recall for high frequency words at learning trial 5 and long delayed free recall, but not at learning trial 1. Patients with DAT showed almost no difference in recall rates for high versus low frequency words, however, the interaction between frequency and diagnosis was not significant. Most interestingly, the interaction between diagnosis and neighborhood size reached statistical significance, with controls having better recall for words with large neighborhood sizes and the opposite being true for persons with DAT.

The results of the study suggest that linguistic recall effects occur in the clinical setting. Given that high frequency words were better recalled than low frequency words across diagnoses, it seems that these effects could be utilized to control the difficulty of wordlist learning tests. The finding that the effects of neighborhood size on recall presented differently for persons with and without DAT revealed a potential problem with existing tests. That is, a linguistic bias favoring certain diagnostic groups, which might decrease the discriminative ability of the instrument. As these results concur with international experimental studies it can be assumed that they are not specific to the German language.

Based on these findings new wordlist learning tests could be constructed according to linguistic criteria that determine the test’s difficulty on the item level, increase the similarity between parallel forms, and aim to avoid a linguistic bias.

5. Applying psycholinguistic evidence to the construction of a new test of verbal memory in late-life cognitive decline: the Auditory Wordlist Learning Test

Authors	Johannes Baltasar Hessler, David Brieber, Johanna Egle, Georg Mandler, Thomas Jahn
Publication status	published in 2017 in *Assessment*, online first.
Copyright	Sage Journals
Location in appendix	C
Individual contribution	The Ph.D.-candidate is the main and first author of the paper, searched the literature, developed the criteria for test construction based on the literature and chapter 4, constructed the test, analyzed the data, and conducted the submission and revision process, all under the supervision of and in accordance with Prof. Thomas Jahn.

Abstract

The manuscript describes the construction of the Auditory Wordlist Learning Test (AWLT) along neurolinguistic criteria that were based on evidence from the literature and the results from chapter 4. The aims in the AWLT's development were to select its words according to their length, frequency, and neighborhood size to control the AWLT's linguistic difficulty on the item level, increase the equivalence between parallel forms as well as between learning and distraction lists, and to avoid interactions with the cognitive status of test-takers as much as possible. This was to be achieved by choosing words that were highly similar with regard to their linguistic characteristics to create homogenous lists. In addition, the AWLT was to be a valuable alternative to the established CVLT and the German wordlist learning test of the CERAD-NTB (CERAD-WL). To achieve this aim, the AWLT was designed to lie between the CVLT and the CERAD-WL with regard to mean linguistic values, learning list length, and number of subtests.

In a first pilot study, the preliminary version of the AWLT in both parallel forms was administered to cognitively healthy persons in the Schuhfried testing center. “Blume” (flower) occurred as a common intrusion (i.e., a word named during recall that was not on the learning list) in the second parallel form of the AWLT. This effect was possibly due to “Blüte” (blossom) being on the learning list and producing the more prototypical “Blume” during recall. Therefore, the second form of the AWLT was adjusted in that “Blüte” was replaced by “Blume”, which was possible with adhering to the construction criteria.

In a second pilot study in the Schuhfried testing center, the final version of the AWLT was again administered to cognitively healthy persons. The AWLT showed good internal consistency as well as test-retest reliability in its core variables, displayed the expected primacy and recency effects in immediate

recall, and showed the expected pattern of scores for learning, recall and recognition in healthy persons.

Multivariate analysis of variance (MANOVA) revealed no difference in mean word length, frequency, and neighborhood size between the four lists of the AWLT (learning and distraction lists of the two parallel forms). Another MANOVA showed no difference in mean test scores on all subtests between parallel forms. Employing repeated measures analysis of variance (RM-ANOVA) on recall rates at learning trials 1 and 4 as well as short and long delayed free recall, a statistically significant main effect of frequency was detected. Also, interactions of subtest with length, frequency, and neighborhood size were statistically significant. Decomposition of the interaction effects with post-hoc *t*-tests and according effect sizes suggested that differences in recall rates of words with high versus low linguistic characteristics (i.e., long versus short, high versus low frequency, large versus small neighborhood) differed only marginally between subtests. The only notable difference was indicated by an advantage for high frequency words at learning trial 4 with a small effect size.

Even though it was not completely possible to avoid effects like the advantage for high frequency words, the AWLT constitutes a reliable wordlist learning tests that is able to assess several aspects of verbal memory. In future studies, the interaction between the AWLT's linguistic profile and the cognitive status of test-takers, as well as its (differential) diagnostic validity, need to be investigated.

6. Linguistic fairness and differential validity of the Auditory Wordlist Learning Test (AWLT) in dementia of the Alzheimer’s type

Authors	Johannes Baltasar Hessler, David Brieber, Johanna Egle, Georg Mandler, Thomas Jahn
Publication status	in print at *Diagnostica*
Location in appendix	D
Individual contribution	The Ph.D.-candidate is the main and first author of the paper, developed the research question, analyzed the data, and conducted the submission process, all under the supervision of and in accordance with Prof. Thomas Jahn.

Abstract

The aims of chapter 6 were twofold. First, we investigated whether the construction principle of linguistic fairness (i.e., no or only small interaction effects between linguistic variables and cognitive status of the test-takers) was reached. Second, we examined the validity of the AWLT's main variables (learning sum, short delayed recall, long delayed recall, and recognition) and the Index Verbal Long-term Memory of the CFD to differentiate between unimpaired older persons and persons with DAT.

The study was based on data from the CFD's norming study and its multicenter validation study. RM-ANOVAs revealed effects of word length, frequency, and neighborhood size on recall rates in the AWLT. Across all levels of subtest and diagnosis, long words were better recalled than short words, frequent words better than infrequent words, and words with large neighborhood sizes better than words with small neighborhood sizes. Further, we found interactions of subtests with word length and word frequency, respectively. Long words were better recalled than short words at the first learning trial but not at the other subtests. Frequent words were better recalled at learning trials 1 and 4 but not at long delayed recall.

The AWLT's main variables and the Index were able to differentiate with large effect sizes between unimpaired persons and patients with DAT. The short delayed recall differentiated best between the groups, followed by the Index, long delayed recall, recognition, and the learning sum. Areas under the curve were all large and statistically significant.

A comparison of these results with the findings from the study in chapter 4, which was conducted according to a very similar design, suggested that the AWLT is linguistically fairer than the CVLT. Even though the AWLT showed linguistic recall effects for all three examined variables, no interaction with

diagnosis was found and effect sizes were mostly smaller than those for the CVLT.

Future studies need to clarify whether the seemingly improved linguistic fairness of the AWLT and its computerized administration also benefit its validity. To this end, the AWLT and other wordlist learning tests, preferably the CVLT, should be administered to the same sample of healthy persons and patients with DAT.

CAVEAT

The article underwent a substantial revision in the peer-review process, which took place after the present dissertation was submitted to the examination board of the faculty. The accepted article differs from the initially submitted article, which is described in the above abstract, discussed in the final part of the dissertation, and enclosed in the appendix, as similar analyses of a larger sample suggested a more complex pattern of linguistic effects in the AWLT. Importantly, all referrals to and conclusions based on the article in the dissertation pertain to the initial, now out-of-date version of the article. The accepted article could not be included in the appendix, as the Hogrefe Verlag, the publisher of *Diagnostica*, has a 12-month embargo on reviewed and accepted papers, which prohibits their provision on online repositories during that time period. Hogrefe only permits the provision of the initially submitted and not yet reviewed version of a manuscript. To access the full text of the accepted paper, please visit https://econtent.hogrefe.com/toc/dia/current. Please note that the date of publication in *Diagnostica* was unclear at the time of printing this dissertation. Hence, we are unable to state as of when the article will be found in the journal's online repository.

7. Discussion

In chapter 7, the findings presented in the previous chapters are synthesized, including a summary of the main results as well as their theoretical and clinical implications. Further, we discuss limitations of the dissertation project and develop perspectives for future research.

7.1 Summary

As described in the introduction, the aims of this dissertation were: (1) To investigate whether interactions between linguistic characteristics and the test takers' cognitive status are present in wordlist learning tests that are used in clinical practice. We found that linguistic recall effects occur in the clinical setting and may impair the validity and fairness of an established instrument. Another aim was (2) to construct a new wordlist learning test according to neurolinguistic criteria based on the evidence from (1) and the available literature. The AWLT is the first wordlist learning test that was developed along strict neurolinguistic criteria. The last aim was (3) to examine whether the newly constructed test is actually linguistically fair and able to assess a range of aspect of verbal memory in persons with and without DAT as well as differentiate between these groups. The AWLT was not free of linguistic interference, however, showed effects that are presumably less problematic than in the CVLT. Also, the AWLT differentiated with large effect sizes between healthy persons and patients with DAT.

In sum, the AWLT is a valid test of verbal memory that can be reliably administered to older persons on a touchscreen. With these features, the AWLT represents a unique instrument in the neuropsychological testing landscape.

7.2 Linguistic recall effects in clinical practice

Linguistic recall effects seem to be relevant not only for experimental studies, but also in the diagnostics of patients with DAT in clinical institutions. In chapters 4 and 6, we found recall advantages for words with high frequency and larger neighborhood sizes, respectively. These findings are paralleled by experimental studies (Jalbert, Neath, & Surprenant, 2011a; MacLeod & Kampe, 1996). While there was no effect of word length in the CVLT, the AWLT showed higher recall rates for longer words. Usually, short words are better recalled than long words in pure lists (Jalbert, Neath, Bireta, & Surprenant, 2011b), as they afford less resources of the limited phonological loop (Baddeley et al., 1975). The results for the AWLT, however, fall in line with a range of equivocal results for mixed lists of both short and long words (Bireta et al., 2006; Jalbert, Neath, Bireta, & Surprenant, 2011b).

Recently, a study employing the Indonesian version of the Hopkins Verbal Learning Test (HVLT; Brandt, 1991) adapted our exact RM-ANOVA method of analyzing recall rates based on word length, frequency, and neighborhood in persons with and without dementia (Grenfell-Essam, Hogervorst, & Rahardjo, 2017). The authors reported a main effect for frequency with higher recall rates for frequent words, as well as an interaction between frequency and subtest. Interestingly, they also found a main effect for neighborhood size, with small neighborhood sizes producing higher recall rates. No interactions between linguistic characteristics and diagnostic group were detected. Together with our findings from the CVLT, these results emphasize that linguistic interference occurs in established verbal memory tests and remains relevant in daily clinical practice.

7.3 Theoretical and clinical implications

The present dissertation adapted a topic from experimental psycholinguistic research and applied it to the clinical field. The findings gained from this transfer highlight that both construction and evaluation of wordlist learning tests potentially benefit from focusing on the individual items' linguistic make-up. Notably, this notion is not new (e.g., see the construction principles of the original CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000), however, to our knowledge, has not been applied to the extent as in the construction of the AWLT. Usually, the psycholinguistic evaluation of verbal memory tests is conducted ex post factum, when the tests are already in use (e.g., Grenfell-Essam et al., 2017; Hessler et al., 2016a). The AWLT, however, was constructed with neurolinguistic criteria established a priori.

Chapters 4 and 5 demonstrated that linguistic recall effects can be seen as both risk and chance. These effects may appear as interference when interacting with the cognitive status of the test-taker, yet, they also provide potential to control the item difficulty and equate parallel test forms. Chapter 6 suggested that the former is presumably not fully preventable, while chapter 5 indicated that the latter is feasible.

Linguistic interference might result from a wordlist that favors one diagnostic group with its linguistic profile. The CVLT, for example, showed an interaction between neighborhood size and cognitive status in recall rates, suggesting that this linguistic variable might introduce bias. On the basis of these findings it could be concluded that in a list including many words with small neighborhood sizes unimpaired persons would still recall a fair number of words, while impaired persons might benefit from the linguistic characteristics. As a result this list would be less able to discriminate between the two groups. In turn, a list showing the opposite linguistic profile might better differentiate between the groups. As this notion would require substantial empirical backup to justify its application to constructing wordlist learning tests, we instead aimed to fully avoid linguistic interactions in the AWLT.

Next to describing linguistic construction principles, the present thesis offers a methodology for evaluating existing wordlist learning tests, which was already adapted by other authors (Grenfell-Essam et al., 2017). Given that chapter 4 demonstrated that these effects occur in a clinical setting with an established instrument, it would be worthwhile to examine other wordlist learning tests with regard to their susceptibility to linguistic interference. Detecting linguistic interference would, of course, not necessarily render a test invalid. However, it would certainly be useful to be able to discuss these aspects in the manuals and inform users about (possible consequences of) the test's linguistic characteristics.

The diagnostic implications of linguistic recall effects were for the first time described in this thesis. Even though chapter 6 demonstrated that no interaction between linguistic characteristics and cognitive status occurred in the AWLT, the proof that this effect is due to the construction principles is not directly given.

7.4 Limitations and future research

The studies presented in this dissertation have strengths and limitations that are discussed in the individual papers. There are, however, a few limitations that apply to the whole of the studies in chapters 4 – 6.

We concluded that the AWLT showed less linguistic interference than the CVLT. This statement needs to be interpreted considering that we applied slightly different methods in analyzing the linguistic recall effects. When examining the CVLT, we used data from a memory clinic and matched persons with or without

DAT based on age, gender, and education. In the evaluation of the AWLT we employed a sample of matched healthy persons and patients with DAT. As a consequence, the controls of the AWLT-study might be cognitively healthier than those from the CVLT-study. Future research should examine the AWLT and other wordlist learning test when administered to the same cognitively unimpaired and impaired persons. This design would on the one hand allow for directly comparing the size of linguistic recall effects, on the other hand, it would be possible to investigate whether the AWLT's construction principles actually increased its validity.

The proposed methodology to examine linguistic recall effects does not allow for stating whether absolute or relative linguistic values are relevant for learning and recall. For example, is it enough for words to be more frequent than other words in the same list to be better recalled or are there absolute values that determine whether a word is frequent or infrequent? Databases like the dlexDB also offer the frequency rank as statistics ("und" is the most common German word). Based on these ranks, words could also be categorized as frequent or infrequent, for example, by dichotomizing the ranks at the middle value. Future research needs to evaluate this alternative and investigate whether similar results as in this dissertation are found when words are linguistically classified based on their absolute and not relative values.

7.5 Conclusion

Linguistic recall effects are relevant in the clinical setting and may potentially influence the validity of established instruments. In order to mitigate this possible influence and at the same time use these effects to gain control over the items, we constructed the AWLT along neurolinguistic criteria. We found the

AWLT to be a valid and reliable wordlist learning tests that potentially shows less linguistic interference than the German CVLT.

Literature

Allen, R. & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, *55*, 64–88.

American Psychiatric Association (2000). *DSM-IV-TR: Diagnostic and statistical manual of mental disorders, text revision (4th ed.)*. Washington, DC: Author.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.) Washington, DC: Author.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575–589.

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: evidence for two processes in episodic recognition performance. *Journal of Memory and Language*, *46*, 199–226.

Bäckman, L., Jones, S., Berger, A.-K., Laukka, E. J., & Small, B. J. (2005). Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. *Neuropsychology*, *19*, 520–531.

Beck, I. R., Schmid, N. S., Berres, M., & Monsch, A. U. (2013). Establishing robust cognitive dimensions for characterization and differentiation of patients with Alzheimer's disease, mild cognitive impairment, frontotemporal dementia and depression. *International Journal of Geriatric Psychiatry*, *29*, 624–634.

Binkau, S., Berwig, M., Jänichen, J., & Gertz, H.-J. (2014). Is the MWT-A suitable for the estimation of premorbid intelligence level? *The Journal of Gerontopsychology and Geriatric Psychiatry*, *27*, 33–39.

Bireta, T. J., Neath, I., & Surprenant, A. M. (2006). The syllable-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review*, *13*, 434–438.

Bondi, M. W., Jak, A. J., Delano-Wood, L., Jacobson, M. W., Delis, D. C., & Salmon, D. P. (2008). Neuropsychological contributions to the early identification of Alzheimer’s disease. *Neuropsychology Review*, *18*, 73–90.

Brandt, J. (1991). The Hopkins Verbal Learning Test: Development of a new memory test with six equivalent forms. *Clinical Neuropsychologist*, *5*, 125–142.

Brébion, G., David, A. S., Bressan, R. A., & Pilowsky, L. S. (2005). Word frequency effects on free recall and recognition in patients with schizophrenia. *Journal of Psychiatric Research*, *39*(2), 215–222.

Brooke, P. & Bullock, R. (1999). Validation of a 6 item cognitive impairment test with a view to primary care usage. *International Journal of Geriatric Psychiatry*, *14*, 936–940.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., B lte, J., & B hl, A. (2011). The Word frequency effect. *Experimental Psychology*, *58*, 412–424.

Butters, M. A., Becker, J. T., Nebes, R. D., Zmuda, M. D., Mulsant, B. H., Pollock, B. G., & Reynolds, C. F., III. (2000). Changes in cognitive functioning following treatment of late-life depression. *American Journal of Psychiatry*, *157*, 1949–1954.

Criss, A. H., Aue, W. R., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language*, *64*, 119–132.

Cunha, C., Guerreiro, M., de Mendonça, A., Oliveira, P. E., & Santana, I. (2012). Serial position effects in Alzheimer's disease, mild cognitive impairment, and normal aging: predictive value for conversion to dementia. *Journal of Clinical and Experimental Neuropsychology*, *34*, 841–852.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *CVLT-II: California verbal learning test: adult version*. San Antonio, TX: The Psychological Corporation.

DeLosh, E. L. & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1136–1146.

Diana, R. A. & Reder, L. M. (2006). The low-frequency encoding disadvantage: word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 805–815.

Egli, S. C., Beck, I. R., Berres, M., Foldi, N. S., Monsch, A. U., & Sollberger, M. (2014). Serial position effects are sensitive predictors of conversion from MCI to Alzheimer's disease dementia. *Alzheimer's & Dementia*, *10*(S), S420–S424.

Erkinjuntti, T., Laaksonen, R., Sulkava, R., Syrjäläinen, R., & Palo, J. (1986). Neuropsychological differentiation between normal aging, Alzheimer's disease and vascular dementia. *Acta Neurologica Scandinavica*, *74*, 393–403.

Foldi, N. S., Brickman, A. M., Schaefer, L. A., & Knutelska, M. E. (2003). Distinct serial position profiles and neuropsychological measures differentiate late life depression from normal aging and Alzheimer's disease. *Psychiatry Research*, *120*, 71–84.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). “Mini-mental state.” A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198.

Fratiglioni, L., De Ronchi, D., & Agüero-Torres, H. (1999). Worldwide prevalence and incidence of dementia. *Drugs and Aging*, *15*, 365–375.

Gainotti, G. (2010). Origins, controversies and recent developments of the MCI construct. *Current Alzheimer Research*, *7*, 271–279.

Gainotti, G., Quaranta, D., Vita, M. G., & Marra, C. (2014). Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease. *Journal of Alzheimer's Disease*, *38*, 481–495.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., ... Winblad, B. (2006). Mild cognitive impairment. *The Lancet*, *367*, 1262–1270.

Geyken, A. (2007). The DWDS corpus: a reference corpus for the German language of the 20th century. *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*, 23–40.

Glanc, G. A., & Greene, R. L. (2007). Orthographic neighborhood size effects in recognition memory. *Memory & Cognition*, *35*, 365–371.

Grenfell-Essam, R., Hogervorst, E., & Rajardjo, T. B. W. (2018). The Hopkins Verbal Learning Test: an in-depth analysis of recall patterns. *Memory*, *26*, 385–405.

Han, S. D., Nguyen, C. P., Stricker, N. H., & Nation, D. A. (2017). Detectable Neuropsychological differences in early preclinical Alzheimer’s disease: a meta-analysis, *Neuropsychology Review*, online first.

Harris, S. J., & Dowson, J. H. (1982). Recall of a 10-word list in the assessment of dementia in the elderly. *The British Journal of Psychiatry*, *141*, 524–527.

Harvey, R. J., Skelton-Robinson, M., & Rossor, M. N. (2003). The prevalence and causes of dementia in people under the age of 65 years. *Journal of Neurology, Neurosurgery & Psychiatry*, *74*, 1206–1209.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*, 10–20.

Hessler, J. B. & Jahn, T. (2017). *Manual Auditiver Wortlisten Lerntest (AWLT)*.

Mödling: Schuhfried GmbH.

Hessler, J. B., Brieber, D., Egle, J., Mandler, G., & Jahn, T. (2017). Applying psycholinguistic evidence to the construction of a new test of verbal memory in late-life cognitive decline. *Assessment*, online first.

Hessler, J. B., Brieber, D., Egle, J., Mandler, G., & Jahn, T. (in print). Lingustic fairness and differential validity of the Auditory Wordlist Learning Test (AWLT) in dementia of the Alzheimer's type. *Diagnostica*.

Hessler, J. B., Fischer, A. M., & Jahn, T. (2016a). Differential linguistic recall effects in the California Verbal Learning Test in healthy aging and Alzheimer's dementia: analysis of routine diagnostic data. *Archives of Clinical Neuropsychology*, *31*, 689–699.

Hessler, J. B., Schäufele, M., Hendlmeier, I., Nora Junge, M., Leonhardt, S., Weber, J., & Bickel, H. (2016b). The 6-Item Cognitive Impairment Test as a bedside screening for dementia in general hospital patients: results of the General Hospital Study (GHoSt). *International Journal of Geriatric Psychiatry*, *32*, 726–733.

Hessler, J., Brönner, M., Etgen, T., Ander, K.-H., Förstl, H., Poppert, H., … Bickel, H. (2014a). Suitability of the 6CIT as a screening test for dementia in primary care patients. *Aging & Mental Health*, *18*, 515–520.

Hessler, J., Jahn, T., Kurz, A., & Bickel, H. (2013). The MWT-B as an estimator of premorbid intelligence in MCI and dementia. *Zeitschrift für Neuropsychologie*, *24*, 129–137.

Hessler, J., Tucha, O., Förstl, H., Mösch, E., & Bickel, H. (2014b). Age-correction of test scores reduces the validity of mild cognitive impairment in predicting progression to dementia. *PLoS ONE*, *9*, e106284–7.

Iadecola, C. (2013). The pathobiology of vascular dementia. *Neuron*, *80*, 844–866.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: effects of word frequency. *Perception & Psychophysics*, *40*, 431–439.

Jahn, T. & Hessler, J. B. (2017). *Handanweisung Kognitive Funktionen Demenz (CFD)*. Mödling: Schuhfried GmbH.

Jahn, T. & Werheid, K. (2015). *Demenzen*. Göttingen: Hogrefe.

Jahn, T., Beitlich, D., Hepp, S., Knecht, R., Köhler, K., Ortner, C., et al. (2013). Drei Sozialformeln zur Schätzung der (prämorbiden) Intelligenzquotienten nach Wechsler. *Zeitschrift Für Neuropsychologie*, *24*, 7–24.

Jahn, T., Theml, T., Diehl, J., Grimmer, T., Heldmann, B., Pohl, C., … Kurz, A. (2004). CERAD-NP und Flexible Battery Approach in der neuropsychologischen Differenzialdiagnostik Demenz versus Depression. *Zeitschrift für Gerontopsychologie & -Psychiatrie*, *17*, 77–95.

Jalbert, A., Neath, I., & Surprenant, A. M. (2011a). Does length or neighborhood size cause the word length effect? *Memory & Cognition*, *39*, 1198–1210.

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. E. M. (2011b). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 338–353.

Jenner, C. & Benke, T. (2002). Neuropsychologische Befunde bei der Frontotemporalen Demenz. *Zeitschrift für Neuropsychologie*, *13*, 161–177.

Jorm, A. F. & Jolley, D. (1998). The incidence of dementia: a meta-analysis. *Neurology*, *51*, 728–733.

Katsumata, Y., Mathews, M., Abner, E. L., Jicha, G. A., Caban-Holt, A., Smith, C. D., … Fardo, D. W. (2015). Assessing the discriminant ability, reliability, and comparability of multiple short forms of the Boston Naming Test in an Alzheimer's disease center cohort. *Dementia and Geriatric Cognitive Disorders*, *39*, 215–227.

Katzman, R., Brown, T., Fuld, P., Peck, A., Schechter, R., & Schimmel, H. (1983). Validation of a short Orientation-Memory-Concentration Test of cognitive impairment. *American Journal of Psychiatry*, *140*, 734–739.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*. *10*, 707–710

Lohnas, L. J. & Kahana, M. J. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1943–1946.

MacLeod, C. M. & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 132–142.

Mitchell, A. J. & Shiri-Feshki, M. (2008). Temporal trends in the long term risk of progression of mild cognitive impairment: a pooled analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, *79*, 1386–1391.

Mitchell, A. J. & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia - meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica*, *119*, 252–265.

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., Fillenbaum, G., et al. (1989). The Consortium to Establish a Registry for Alzheimer"s Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, *39*, 1159–1165.

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., … Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*, 695–699.

Niemann, H., Sturm, W., Thöne-Otto, A. I., & Willmes, K. (2008). *California Verbal Learning Test. Deutsche Adaption*. Frankfurt am Main: Pearson Assessment & Information GmbH.

Ozubko, J. D. & Joordens, S. (2007). The mixed truth about frequency effects on free recall: effects of study list composition. *Psychonomic Bulletin & Review*, *14*, 871–876.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*, 1.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., & Ferri, C. P. (2013). The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's & Dementia*, *9*, 63–75.

Raina, P., Santaguida, P., Ismaila, A., Patterson, C., Cowan, D., Levine, M., et al. (2008). Effectiveness of cholinesterase inhibitors and memantine for treating dementia: evidence review for a clinical practice guideline. *Annals of Internal Medicine*, *148*, 379–397.

Reitz, C., Brayne, C., & Mayeux, R. (2011). Epidemiology of Alzheimer disease. *Nature Publishing Group*, *7*, 137–152.

Riedl, L., Mackenzie, I., Förstl, H., Kurz, A., & Diehl-Schmid, J. (2014). Frontotemporal lobar degeneration: current perspectives. *Neuropsychiatric Disease and Treatment*, *10*, 297–14.

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1019–1034.

Rubin, D. C. & Friendly, M. (1986). Predicting which words get recalled: measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns. *Memory & Cognition*, *14*, 79–94.

Shacter, D. L., & Wagner, A. D. (2013). Learning and Memory. In E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, & A. J. Hudspeth (Eds.), *Principles of Neural Science* (5 ed., pp. 1441–1460). New York: McGraw-Hill.

Soo-ampon, S., Wongwitdecha, N., Plasen, S., Hindmarch, I., & Boyle, J. (2004). Effects of word frequency on recall memory following lorazepam, alcohol, and lorazepam alcohol interaction in healthy volunteers. *Psychopharmacology*, *176*, 420–425.

Tahlmann, B., & Monsch, A. U. (1997). CERAD - The Consortium to Establish a Registry for Alzheimer's Disease. Autorisierte deutsche Fassung. Basel: Memory Clinic Basel.

Van Overschelde, J. P. (2002). The influence of word frequency on recency effects in directed free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 611–615.

Vernooij-Dassen, M., Vasse, E., Zuidema, S., Cohen-Mansfield, J., & Moyle, W. (2010). Psychosocial interventions for dementia patients in long-term care. *International Psychogeriatrics*, *22*(07), 1121–1128.

Wilson, R. S., Bacon, L. D., Fox, J. H., Kramer, R. L., & Kaszniak, A. W. (2008). Word frequency effect and recognition memory in dementia of the Alzheimer type. *Journal of Clinical Neuropsychology*, *5*, 97–104.

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.-O., et al. (2004). Mild cognitive impairment – beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine*, *256*, 240 – 246.

Wolk, D. A., & Dickerson, B. C. (2011). Fractionating verbal episodic memory in Alzheimer's disease. *NeuroImage*, *54*, 1530–1539.

World Health Organization (1993). *The ICD-10 classification of mental and* *behavioural disorders*. Geneva: Author.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: a review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.

Zakzanis, K. K., Leach, L., & Kaplan, E. (1999). Dementia of the Alzheimer's type. In *Neuropsychological differential diagnosis* (pp. 33–33). Leiden: Swets & Zeitlinger.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*, 1–29.

Appendix

Appendix A

TABLE A. Definitions of dementia according to ICD-10, DSM-IV, and DSM-5.

ICD-10	DSM-IV-TR	DSM-5
G1.1 A decline in memory, which is most evident in the learning of new information, although in more severe cases, the recall of previously learned information may be also affected. The impairment applies to both verbal and non-verbal material. The decline should be objectively verified by obtaining a reliable history from an informant, supplemented, if possible, by neuropsychological tests or quantified cognitive assessments.	A. The development of multiple cognitive deficits manifested by both memory impairment (impaired ability to learn new information or to recall previously learned information) and ...	A. Evidence of significant cognitive decline from a previous level of performance in one or more cognitive domains: Learning and memory, Language, Executive function, Complex attention, Perceptual-motor, Social cognition The evidence of decline is based on: Concern of the individual, a knowledgeable informant, or the clinician that there has been a significant decline in cognitive function; and a substantial impairment in cognitive performance, preferably documented by standardized neuropsychological testing or, in its absence, another quantified clinical assessment.
G1.2 A decline in other cognitive abilities characterized by deterioration in judgment and thinking, such as planning and organizing, and in the general processing of information. Evidence for this should be obtained when possible from interviewing an informant, supplemented, if possible, by neuropsychological tests or quantified objective assessments. Deterioration from a previously higher level of performance should be established.	… one (or more) of the following cognitive disturbances: (a) aphasia (language disturbance); (b) apraxia (impaired ability to carry out motor activities despite intact motor function); (c) agnosia (failure to recognize or identify objects despite intact sensory function); (d) disturbance in executive functioning (i.e., planning, organizing, sequencing, abstracting)	
Addition to G1. The decline of cognitive abilities causes impairment in daily activities.	B. The cognitive deficits in Criteria A1 and A2 each cause significant impairment in social or occupational functioning and represent a significant decline from a previous level of functioning.	B. The cognitive deficits interfere with independence in everyday activities. At a minimum, assistance should be required with complex instrumental activities of daily living, such as paying bills or managing medications.
G2. Preserved awareness of the environment during a period of time long enough to enable the unequivocal demonstration of G1. When there are superimposed episodes of delirium the diagnosis of dementia should be deferred.	E. The deficits do not occur exclusively during the course of a delirium.	C. The cognitive deficits do not occur exclusively in the context of a delirium.

	F. The disturbance is not better accounted for by another Axis I disorder (e.g., Major Depressive Episode, Schizophrenia).	D. The cognitive deficits are not better explained by another mental disorder (e.g., major depressive disorder, schizophrenia).	
G3. A decline in emotional control or motivation, or a change in social behavior, manifest as at least one of the following: (1) emotional lability; (2) irritability; (3) apathy; (4) coarsening of social behavior.			
G4. For a confident clinical diagnosis, G1 should have been present for at least six months; if the period since the manifest onset is shorter, the diagnosis can only be tentative.	C. The course is characterized by gradual onset and continuing cognitive decline.		
	D. The cognitive deficits in Criteria A1 and A2 are not due to any of the following: (1) other central nervous system conditions that cause progressive deficits in memory and cognition (e.g., cerebrovascular disease, Parkinson's disease, Huntington's disease, subdural hematoma, normal-pressure hydrocephalus, brain tumor) (2) systemic conditions that are known to cause dementia (e.g., hypothyroidism, vitamin B or folic acid deficiency, niacin deficiency, hypercalcemia, neurosyphilis, HIV infection) (3) substance-induced conditions		
		Differentiation between minor and major neurocognitive disorder should be based on standardized neuropsychological testing.	
		Minor. Test performance between 1 and 2 standard deviations below the group mean.	Mayor: Test performance lower than 2 standard deviations from the group mean.

Appendix B

Differential linguistic recall effects in the California Verbal Learning Test in healthy aging and Alzheimer's dementia: analysis of routine diagnostic data

OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 31 (2016) 689–699

Archives of CLINICAL NEUROPSYCHOLOGY

Differential Linguistic Recall Effects in the California Verbal Learning Test in Healthy Aging and Alzheimer's Dementia: Analysis of Routine Diagnostic Data

Johannes Baltasar Hessler*, Alina Maria Fischer, Thomas Jahn

Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

*Corresponding author at: Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technical University of Munich, Ismaninger Strasse 22, 81675 Munich, Germany. Tel.: +49 89 4140 6183; fax: +49 89 4140 6379
E-mail address: johannes.hessler@tum.de (J.B. Hessler).

Accepted 14 June 2016

Abstract

Objective: Psycholinguistic evidence suggests that certain word characteristics might influence recall rates in word-list learning tests. These effects were investigated in the German California Verbal Learning Test (CVLT-G) in a clinical setting.
Method: Subjects were memory clinic patients without cognitive diagnosis ($N = 45$) and with dementia of the Alzheimer type (DAT) ($N = 48$) matched for age, sex, depressive symptoms, and education. The CVLT-G's words were analyzed with regard to length, frequency, and neighborhood size and dichotomized into low and high value groups. For each linguistic variable, a 2 (*diagnosis*: control vs. DAT) × 3 (*time*: Trial 1 vs. Trial 5 vs. Long Delay Free Recall) × 2 (*linguistic*: low vs. high) repeated measures analysis of variance (RM-ANOVA) was conducted.
Results: RM-ANOVAs revealed a main effect for *frequency*, $F(1,91) = 21.03$, $p < 0.001$, and interactions between *time* and *frequency*, $F(1.97,179.09) = 5.18$, $p = 0.007$, and *diagnosis* and *neighborhood*, $F(1.77,161.23) = 13.60$, $p < 0.001$. High-frequency words were better recalled at Trial 5 (Cohen's $d = 0.37$) and long delay*d* free recall ($d = 0.16$) and learning from Trials 1 to 5 was better for high-frequency words ($d = 0.39$). Controls recalled large neighborhood words better whereas the opposite was true for persons with DAT ($d = 0.76$).
Conclusion: Frequency and neighborhood size seem to influence learning and retention in the CVLT-G with neighborhood size producing opposed effects for persons with and without DAT. These results are in line with international experimental studies and likely not specific to the German language. Potential diagnostic implications and possibilities for test construction and interpretation are discussed.

Keywords: Word Frequency; Neighborhood Size; Word Length; CVLT; Verbal Memory

Introduction

Assessing the memory of verbally presented material is an important part in the diagnostics of dementia of the Alzheimer type (DAT). Verbal memory deficits usually emerge early in the course of Alzheimer's disease (Bäckman, Jones, Berger, Laukka & Small, 2005) and are thought to reflect the associated neuropathologic changes (Bondi et al., 2008). Differential patterns of verbal memory recall and recognition also support the discrimination of DAT from depression (Foldi, Brickman, Schaefer & Knutelska, 2003; Wright & Persad, 2007) and frontotemporal dementia (Diehl & Kurz, 2002), and verbal memory impairment can predict progression from mild cognitive impairment to dementia (Gainotti, Quaranta, Vita & Marra, 2014).

The California Verbal Learning Test

The California Verbal Learning Test in its second version (CVLT-II; Delis, Kramer, Kaplan & Ober, 2000) is frequently and internationally employed for the assessment of verbal memory in the context of DAT.

doi:10.1093/arclin/acw038 Advance Access publication on 14 July 2016

The words of the CVLT-II are concrete and are equally distributed across four semantic categories. The CVLT-II's words were chosen with regard to their prototypicality of the respective semantic category and their frequency (Delis et al., 2000). The authors excluded the four most prototypical words of each category in order to prevent inflated recall scores in patients with confabulation tendencies, who often report prototypical intrusions. The mean word frequency was increased compared to the previous version to compile a list that is easier to learn and recall. In the construction of the German CVLT (CVLT-G; Niemann, Sturm, Thöne-Otto & Willmes, 2008), which is examined in this study, words of two high-frequency categories (clothes, kitchenware) and two low-frequency categories (vegetables, fish) were selected to avoid floor and ceiling effects, respectively. Parallel to the original version, the four most prototypical words of each category were excluded based on available norms (Mannhaupt, 1983). The CVLT-II and the CVLT-G are equal in their structure and scoring. Whereas the original CVLT already exists in its second version, the CVLT-G exists in only one version, which had the CVLT-II as its model.

Evidence from psycholinguistic studies suggests an association of certain word characteristics with recall and recognition probability (Rubin & Friendly, 1986). As its linguistic profile was not controlled in the construction, it is unknown whether the CVLT-G is subject to psycholinguistic recall effects and how these effects might present in the clinical setting.

Linguistic Verbal Memory Effects

Orthographic length, word frequency, and neighborhood size appear to be particularly relevant for performance in word-list learning paradigms. Orthographic length can be operationalized by the number of syllables (Jalbert, Neath, Bireta & Surprenant, 2011b). The word-length effect describes the fairly robust finding that pure lists of short words are better recalled than pure lists of long words (Jalbert et al., 2011b). CVLT-II and CVLT-G are mixed lists that contain both short and long words. Whereas some studies found no difference in the recall of short and long words in mixed lists (Bireta, Neath & Surprenant, 2006), others found an advantage for either long (Katkov, Romani & Tsodyks, 2014) or short words (Hulme, Suprenant, Bireta, Stuart & Neath, 2004).

Word frequency describes the commonness of a given word in its language (Criss, Aue & Smith, 2011) and is usually determined on the basis of linguistic databases. The "word frequency mirror effect" in pure lists describes the concurrence of a recall advantage for high-frequency words with a recognition advantage for low-frequency words (MacLeod & Kampe, 1996). In mixed lists, no recall advantage seems to exist, which could be explained by a U-shaped association between frequency and recall probability (Lohnas & Kahana, 2013). In lists that randomly alternated high- and low-frequency words, low-frequency words were better recalled (Ozubko & Joordens, 2007). As word frequency was not further considered in the construction of the CVLT-G, we assume it belongs to the randomly mixed lists of both frequent and infrequent words.

Following Levenshtein's algorithm (Levenshtein, 1966), neighborhood size is defined by the number of words that can be created by either replacing, adding, or deleting a single letter in a word at a time. Large neighborhood sizes are usually related to a higher recall probability (Jalbert et al., 2011b; Roodenrys, Hulme, Lethbridge, Hinton & Nimmo, 2002). In recognition tasks, small neighborhood sizes have been shown to produce more correct identifications and less false alarms (Glanc & Greene, 2007). Presumably, the neighborhood sizes vary unsystematically in the CVLT-G and CVLT-II.

Linguistic Recall Effects in Persons with DAT

Importantly, these findings were established with cognitively healthy participants in experimental settings. Persons with and without DAT certainly show a quantitative difference in verbal memory, but there are few investigations of qualitative differences. So far, linguistic effects have only been examined in recognition, but not in recall. The usually found recognition advantage for low-frequency words was eliminated in this group, whereas the hit rate for high-frequency words remained similar to that of healthy controls (Balota, Burgess, Cortese & Adams, 2002; Wilson, Bacon, Fox & Kramer, 1983).

This Study

The aim of this study was to map the linguistic profile of the CVLT-G with regard to word length, word frequency, and neighborhood size. Further, we investigated whether these linguistic characteristics affect learning and retention performance in the CVLT-G in memory clinic patients with and without DAT. We hypothesized that the CVLT-G's words would show considerable linguistic variability. Further, we expected shortness, high frequency, and large neighborhood size to increase the words' probability of being recalled. As there is limited previous evidence, the investigation of the interaction between DAT and the linguistic factors was exploratory.

Method

Participants and Procedure

Participants were retrospectively identified from archived patient records of the Department of Psychiatry and Psychotherapy at the Klinikum rechts der Isar of the Technical University of Munich. These patients were initially examined at the department's Center for Cognitive Disorders with suspected neurodegenerative disease and underwent cognitive testing at the department's neuropsychological unit between January 1998 and August 2008.

The study protocol was approved by the ethics committee of the Faculty of Medicine of the Technical University of Munich (approval number 81/16 S). Only anonymized routine diagnostic data were analyzed and no analyses according to age, sex, education, place of residence, year of examination, or any other variable that might be used to identify individual patients were performed.

For the purpose of this study, we selected two groups of participants that were matched for age, sex, education, and depressive symptoms. First, patients for whom no cognitive impairment could be quantified, that is, who had not received diagnoses of cognitive impairment or of any other psychiatric or neurological condition. Second, persons who had received a diagnosis of DAT according to ICD-10 (F00.x) after exclusion of other neurological or psychiatric conditions. No further exclusion criteria were applied in both groups. All diagnoses were based on a thorough neuropsychological and physiological examination. Persons with DAT were all diagnosed for the first time in the course of the disease, which suggests that the severity of DAT was rather mild or moderate.

At the memory clinic, the neuropsychological test battery of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD; Morris, Heyman, Mohs & Hughes, 1989; Thalmann & Monsch, 1997), structural and/or functional neuroimaging, as well as cerebrospinal fluid diagnostics were conducted. At the neuropsychological unit, a flexible battery approach (Jahn, Theml, Diehl & Grimmer, 2004) was employed that included the CVLT-II and Beck Depression Inventory (BDI; Beck, Steer & Brown, 1996; Hautzinger, Keller & Kühner, 2006). CVLT-G and CVLT-II are similar in structure and scoring. The German version was not directly translated by its authors but designed to be as close as possible to the CVLT-II (Niemann et al., 2008).

Eligible participants were identified by means of an inventory of the neuropsychological unit listing demographic data, and, among others the cognitive discharge diagnoses. We retrieved the respective patient files from clinic archives, extracted the CVLT-G score sheets, and transcribed the raw data into an Excel-file. Each word was treated as a single binary variable indicating whether the respective patient recalled the word or not. We analyzed linguistic recall effects at Trials 1 and 5 of the learning phase, as well as at the Long Delay Free Recall as indicator of retention.

Linguistic Analysis

Word length was operationalized as the number of a word's syllables. Word frequency and neighborhood size were determined with the dlexDB (Heister et al., 2011), a German lexical database that uses the text corpus of the Digitales Wörterbuch der deutschen Sprache (DWDS; Geyken, 2007). The DWDS comprises 2,224,542 different words extracted from prose, newspaper articles, functional texts, and transcribed spoken language from the whole 20th century. The dlexDB is accessible online (www.dlexdb.de) and free of charge.

The option "annotated normalized type frequency" was selected as a parameter indicating word frequency. Types represent the general form of a word, regardless of its different semantic meanings. "Annotated" means that frequency values are separately listed for words that are orthographically similar but represent different parts of speech. For example, the German word "Frage" means "question", whereas "frage" is the first person singular verb of "to ask". By doing so we were able to extract the frequency of the nouns from the database. "Normalization" is a standardization method that indicates a word's frequency per one million tokens in the corpus. A token represents a word's specific occurrence in the corpus. For example, the series AABBCCDD includes four types (A, B, C, D) and eight tokens.

"Normalized neighborhood" according to Levenshtein's algorithm (Levenshtein, 1966) was selected as an indicator for neighborhood size. This method counts each word as a neighbor that can be derived from the original word by either adding, removing, or replacing a single letter. For example, the words "hause" (first person singular of "to dwell") "aus" ("out"), and "Maus" ("mouse") are neighbors of "Haus" ("house"). Per neighbor only one of these operations can be performed.

Statistical Analysis

Usually, psycholinguistic studies are conducted in an experimental setting with high control over word selection (e.g., three syllable words versus five syllable words) and list composition (pure, randomly mixed, or systematically mixed). Because we analyzed a preexisting list, this method was not applicable. Instead, we dichotomized the list using the median of each linguistic characteristic as cut-off. Specifically, for each linguistic characteristic, the 16 words of the CVLT-G were divided into two groups (i.e., short vs. long; high vs. low frequency; small vs. large neighborhood size). As the median splits likely produce lists of unequal length, it was not possible to compare the number of recalled words in each linguistic group. Instead, the percentage of correctly recalled words within each linguistic group at Trials 1 and 5 of the immediate recall and at Long Delay Free Recall was calculated to assess linguistic recall effects on learning and retention.

The CVLT-G word list can be divided into a primacy area (first four words), a middle area (middle eight), and a recency area (last four). Words in the primacy and recency areas are usually better recalled than words in the middle area (Murdock, 1962). These primacy and recency effects could confound the results of the subsequent analyses, if the linguistic characteristics were not evenly distributed across the areas of the CVLT-G. In order to rule out this confounding factor, we performed a series of χ^2-tests on the distribution of each of the dichotomized linguistic variables across the three areas.

For each linguistic characteristic, we performed a separate 2 (*diagnosis*: control vs. DAT) × 3 (*time*: Trial 1 vs. Trial 5 vs. Long Delay Free Recall) × 2 (*linguistic*: low vs. high) repeated measures analysis of variance (RM-ANOVA) of recall rates with *time* and *linguistic* as within-subjects factors and *diagnosis* as between-subject factor. Each RM-ANOVA analyzes the same variance for *diagnosis* and *time*. Therefore, we corrected the *p*-values for the main effects of diagnosis and time and their interaction with the Bonferroni method (multiplied with 3). As it can be expected that recall rates differ between diagnostic groups and across time, *post hoc t*-tests were only calculated for significant interaction effects involving linguistic factors.

Multivariate analysis of variance (MANOVA) was employed to compare differences in recognition performance between persons with and without DAT in relation to the words' linguistic characteristics. Dependent variables were the differences in the percentage of correctly recognized words (true positives) in the group of words that were either low or high on a given linguistic variable (e.g., short vs. long words). A 2 (*diagnosis*: control vs. DAT) × 3 (*linguistic*: difference short vs. long length, difference low vs. high frequency, difference small vs. large neighborhood size) MANOVA of these differences in recognition rates was then performed.

As we aimed to match the diagnostic groups with regard to their size, the number of the total sample was determined by the number of patients who were discharged without a diagnosis of cognitive impairment. From patient records, we knew that this number was around 45 for the indicated period of time. Therefore, this value was employed as the size for the smallest group in the power analysis that was conducted with GLIMMPSE (http://glimmpse.samplesizeshop.org/#/). Following the guidelines by Guo, Logan, Glueck and Muller (2013), a power of 0.82 was calculated to detect any effects with the chosen design and the expected sample size.

Data analysis was performed with SPSS 23 for Microsoft Windows. For *post hoc t*-tests 95% confidence intervals of the mean difference and Cohen's *d* as a measure of effect size are reported. *P*-values were adjusted for multiple comparisons with the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

Results

Forty-seven persons without a cognitive diagnosis were identified across the whole recruitment period. Fifty persons with DAT were matched with regard to age, sex, and education. In either group, two did not complete the CVLT-G and were excluded. Ninety-three (50 male, 53.8%) participants with a mean age of 63.38 years ($SD = 8.38$) were included in the analyses. Further descriptive statistics according to diagnosis are shown in Table 1. The results of the CERAD-NTB suggest that the persons with DAT were in mild-to-moderate stages of dementia. For example, their mean MMSE score was around 23.

The CVLT-G displayed considerable linguistic heterogeneity (Table 2). The number of syllables ranged from 1 to 4 (median = 2), normalized word frequency from 0.06 to 5.28 (median = 1.17), and normalized neighborhood size from 0 to 18.41 (median = 2.78). The respective median was used to group the words for each linguistic variable into high and low values. χ^2-tests for the distribution of the dichotomized linguistic variables across primacy, middle, and recency areas suggested no differences for word length, $\chi^2(2, N = 45) = 0.29$, $p = 0.659$, frequency, $\chi^2(2, N = 45) = 0.25$, $p = 0.938$, and neighborhood size, $\chi^2(2, N = 45) = 1.50$, $p = 0.732$.

Figure 1 displays the recall rates of the words according to diagnosis and linguistic groups.

The degrees of freedom for *time* were corrected with the Greenhouse–Geisser formula due to a violation of sphericity in all subsequent RM-ANOVAs.

Word length did not influence recall rates. Aggregated across *time* and *length*, controls had higher recall rates ($M = 60.82$, $SD = 13.09$) than persons with DAT ($M = 23.21$, $SD = 13.09$), $F(1,91) = 191.62$, $p < 0.001$, partial $\eta^2 = 0.68$. Aggregated across *diagnosis* and *length*, recall rates increased from Trial 1 ($M = 30.40$, $SD = 11.62$) to Trial 5 ($M = 55.58$, $SD = 15.12$) and decreased again to Long Delay Free Recall ($M = 40.06$, $SD = 19.59$), $F(1.86,169.15) = 128.98$, $p < 0.001$, partial

Table 1. Sample descriptive statistics as well as results of the German California Verbal Learning Test (CVLT-G), the Consortium to Establish a Registry for Alzheimer's Disease neuropsychological test battery (CERAD-NTB), and Beck Depression Inventory (BDI) for controls and persons with dementia of the Alzheimer type (DAT)

Variable	Control, $N = 45$	DAT, $N = 48$	p[a]
Age; M (SD)	62.38 (8.08)	65.31 (8.63)	0.118
Female; N (%)	21 (46.7)	22 (45.8)	0.300
Education; N (%)			0.799
Secondary modern school	10 (22.2)	14 (29.2)	
Higher schools	35 (77.8)	34 (70.8)	
BDI; M (SD)	7.69 (5.12)	9.69 (9.27)	0.401
CVLT-G score; M (SD)			
Trial 1	6.69 (2.26)	2.90 (1.12)	<0.001
Trial 5	12.09 (2.27)	5.79 (2.33)	<0.001
Long Delay Free Recall	10.53 (3.28)	2.33 (2.85)	<0.001
CERAD-NTB raw scores; M (SD)			
Semantic word fluency	21.87 (7.07)	12.06 (5.13)	<0.001
Boston Naming Test	14.51 (0.89)	12.94 (1.96)	<0.001
Mini Mental Status Test	28.84 (1.40)	23.23 (3.63)	<0.001
Word-list learning sum	20.56 (3.65)	13.04 (4.21)	<0.001
Word-list delayed recall	7.04 (1.81)	2.35 (2.13)	<0.001
Word-list discriminability	97.22 (6.08)	83.67 (12.03)	<0.001
Visuoconstruction	10.42 (1.03)	8.59 (2.15)	<0.001
Visuoconstruction delayed	8.76 (2.81)	3.53 (3.11)	<0.001

[a]Independent t-tests for continuous variables and χ^2-tests for categorical variables. Percentages apply per column. M = mean. SD = standard deviation.

Table 2. Number of syllables, type frequencies, and neighborhood sizes for the words of the German California Verbal Learning Test. Results of the linguistic analysis

Word	English translation	Syllables	Type frequency[a]	Neighborhood size[a]
Gurke	Cucumber	**2**	1.86	7.28
Toaster	Toaster	**2**	**0.14**	3.42
Schal	Scarf	**1**	2.85	17.12
Kabeljau	Codfish	3	**0.50**	**0.43**
Dosenöffner	Can opener	4	**0.02**	**0**
Lachs	Salmon	**1**	**1.17**	14.98
Krawatte	Tie	3	4.85	**1.28**
Porree	Leeks	**2**	**0.86**	4.28
Makrele	Mackerel	3	**0.24**	**0.43**
Quirl	Beater	**1**	**0.27**	**2.14**
Zwiebeln	Onion	**2**	5.28	**2.14**
Bluse	Blouse	**2**	5.24	6.85
Rotbarsch	Redfish	**2**	**0.06**	**0.86**
Sieb	Sieve	**1**	2.35	18.41
Kohlrabi	Kohlrabi	3	**1.17**	**0.86**
Socken	Socks	**2**	3.29	11.99
Median		2	1.17	2.78

[a]Normalized, that is, per million words in the corpus. **Boldface** indicates values ≤ median.

$\eta^2 = 0.59$. Aggregated across *diagnosis* and *time*, there were no differences in recall rates between short and long words, $F(1,91) < 0.01$, $p = 0.995$, partial $\eta^2 < 0.01$. The interaction between *diagnosis* and *time* was significant, $F(1.86,169.15) = 32.20$, $p < 0.001$, partial $\eta^2 = 0.26$. The interactions between *diagnosis* and *length*, $F(1,91) = 0.66$, $p = 0.417$, partial $\eta^2 = 0.01$, as well as *time* and *length*, $F(1.95,177.31) = 1.26$, $p = 0.287$, partial $\eta^2 = 0.01$, were not significant. The interaction between *diagnosis, time*, and *length*, $F(1.95,177.31) = 2.60$, $p = 0.078$, partial $\eta^2 = 0.03$ was marginally significant.

Frequency influenced recall rates in the same manner across diagnostic groups but differently over time. Aggregated across *time* and *frequency*, controls had higher recall rates ($M = 61.57$, $SD = 12.70$) than persons with DAT ($M = 23.18$, $SD = 12.70$), $F(1,91) = 212.31$, $p < 0.001$, partial $\eta^2 = 0.70$. Aggregated across *diagnosis* and *frequency*, recall rates increased from Trial 1 ($M = 30.06$, $SD = 11.03$) to Trial 5 ($M = 56.49$, $SD = 14.18$) and decreased at Long Delay Free Recall ($M = 40.57$, $SD = 19.35$), $F(1.77,160.84) = 154.81$, $p < 0.001$, partial $\eta^2 = 0.63$. Aggregated across *diagnosis* and

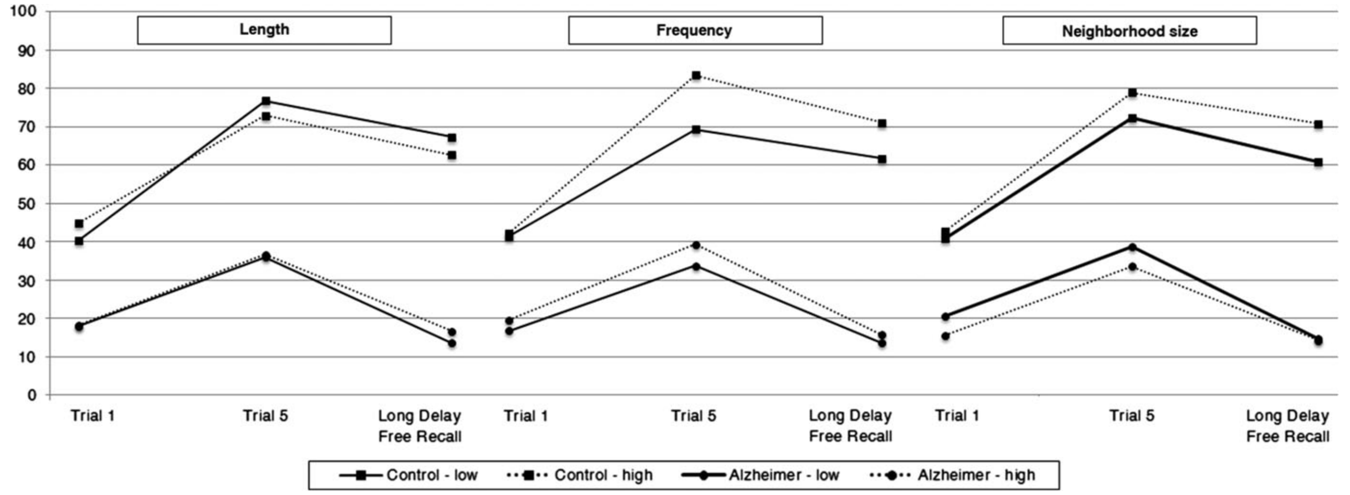


Fig. 1. Mean recall rates at Trial 1, Trial 5, and Long Delay Free Recall according to diagnosis as well as word length, frequency, and neighborhood size.

time, high-frequency words were better recalled (M = 44.60, SD = 24.60) than low-frequency words (M = 38.91, SD = 23.10), F(1,91) = 21.03, $p < 0.001$, partial $\eta^2 = 0.19$. The interaction between *diagnosis* and *time* was significant, F(1.77,160.84) = 43.59, $p < 0.001$, partial $\eta^2 = 0.32$. The interaction between *time* and *frequency* indicated different recall rates for high- and low-frequency words at different times, F(1.97,179.09) = 5.18, $p = 0.007$, partial $\eta^2 = 0.05$. The interactions between *diagnosis* and *frequency, F*(1,91) = 3.38, $p = 0.069$, partial $\eta^2 = 0.04$, as well as *diagnosis*, *time,* and *frequency* (F(1.97,179.09) = 2.67, $p = 0.073$, partial $\eta^2 = 0.03$) were marginally significant.

Neighborhood size affected recall differently for controls and persons with DAT, but these differential patterns were consistent over time. Aggregated across *time* and *neighborhood*, controls had higher recall rates (M = 61.07, SD = 12.77) than persons with DAT (M = 22.96, SD = 12.77), F(1,91) = 206.95, $p < 0.001$, partial $\eta^2 = 0.70$. Aggregated across diagnosis and neighborhood, recall rates increased from Trial 1 (M = 29.95, SD = 11.01) to Trial 5 (M = 55.88, SD = 14.39) and decreased at Long Delay Free Recall (M = 40.21, SD = 19.15), F(1.77,161.23) = 154.14, $p < 0.001$, partial $\eta^2 = 0.63$. Aggregated across *time* and *diagnosis*, there was no difference in recall rates between words with small and large neighborhood sizes, F(1,91) = 1.00, $p = 0.320$, partial $\eta^2 = 0.01$. The interactions between *diagnosis* and *time*, F(1.77,161.23) = 43.14, $p < 0.001$, partial $\eta^2 = 0.32$, and between *diagnosis* and *neighborhood, F*(1,91) = 13.60, $p < 0.001$, partial $\eta^2 = 0.13$, were significant. The interaction between *time* and *neighborhood* was marginally significant, F(1.94,176.16) = 2.38, $p = 0.097$, partial $\eta^2 = 0.03$. The interaction between *diagnosis*, *time*, and *neighborhood* was not significant*, F*(1.94,176.16) = 0.40, $p = 0.668$, partial $\eta^2 < 0.01$. Table 3 displays the results of the significant *post hoc* tests for effects involving linguistic characteristics.

In combination with Fig. 1, the results suggest a recall advantage for high-frequency words. Also, the recall rates for low- and high-frequency words showed a differential effect over time. Whereas there was almost no difference at Trial 1, the high-frequency words were better recalled at Trial 5 and Long Delay Free Recall. All effects were more pronounced in the control group, the three-way interaction between *diagnosis*, *time*, and *frequency*, however, was only marginally significant.

Figure 1 and the results from Table 3 show a clear and statistically significant interaction effect between *diagnosis* and *neighborhood size*. At all times, the controls had better recall rates for words with large compared to small neighborhood sizes. For persons with DAT, this effect was reversed with an advantage for words with small neighborhood sizes. The differences in mean recall rates between large and small neighborhood size words were statistically significantly different for controls and persons with DAT with a large effect size. Whereas the difference increased over time for controls, persons with DAT showed no differences in recall rates at Long Delay Free Recall anymore. The three-way interaction between *diagnosis*, *time*, and *neighborhood*, however, was not statistically significant.

MANOVA revealed that persons with and without DAT did not differ in the difference scores for correct recognition rates in relation to length, frequency, and neighborhood size, Wilk's $\lambda = 0.94$, $p = 0.127$, partial $\eta^2 = 0.06$. Though the difference scores were not significantly different, there were some interesting trends. Whereas persons without DAT showed better recognition performance for words with high frequency (mean difference low–high = −3.10, SD = 12.77), there was almost no difference in recognition rates in persons with DAT (mean difference = 0.10, SD = 16.90). Similarly, persons without DAT showed better recognition rates for words with large neighborhood sizes (mean difference small–large neighborhood size = −2.78, SD = 12.46), whereas persons with DAT showed the opposite pattern (mean difference = 1.82, SD = 16.10).

Table 3. *Post hoc t*-tests and mean recall rates for statistically significant interaction effects involving linguistic factors

Linguistic recall effects	% recalled			*T* (*df*)	Cohen's *d*
	M (*SD*)	Difference *M* (*SD*)	Difference 95% CI		
Frequency					
Interaction with time					
Trial 1					
Low	28.79 (18.33)	−1.78 (17.58)	−5.40; 1.84	−0.98 (92)	0.10
High	30.57 (18.43)				
Trial 5					
Low	51.02 (25.39)	−9.66 (19.60)	−13.70; −5.62	−4.75* (92)	0.37
High	60.68 (27.35)				
Long Delay Free Recall					
Low	36.92 (31.13)	−5.63 (18.73)	−9.49; −1.78	−2.90* (92)	0.16
High	42.55 (35.99)				
Low: Trial 1–Trial 5	−22.22 (17.49)	7.88 (25.94)	2.54; 13.22	2.93* (92)	−0.39
High: Trial 1–Trial 5	−30.11 (22.24)				
Low: Trial 1–LDFR	−8.13 (21.62)	3.86 (24.15)	−1.12; 8.83	1.54 (92)	−0.15
High: Trial 1–LDFR	−11.98 (28.20)				
Low: Trial 5–LDFR	14.10 (16.64)	−4.03 (22.80)	−8.72; 0.67	−1.70 (92)	0.23
High: Trial 5–LDFR	18.13 (20.33)				
Neighborhood size					
Interaction with diagnosis[a]					
Control					
Small	57.96 (16.13)	−6.20 (14.01)	−10.41; −2.0	−2.97* (92)	0.39
Large	64.17 (15.85)				
DAT					
Small	24.74 (12.67)	3.56 (11.47)	0.22; 6.89	2.15* (92)	−0.29
Large	21.18 (12.21)				
Control: small–large	−6.20 (14.01)	−9.76 (2.65)	−15.02; −4.50	−3.69* (91)	−0.76
DAT: small–large	3.56 (11.47)				

*Significant according to the Benjamini–Hochberg method for the adjustment of multiple comparisons. 95% CI = 95% confidence interval. DAT = dementia of the Alzheimer type.

[a] Comparison of marginal means across time.

Discussion

This study examined linguistic recall effects in the German CVLT in an clinical setting. The CVLT-G's words showed considerable heterogeneity with regard to length, frequency, and neighborhood size. High-frequency words were better recalled than low-frequency words in general and learning of high-frequency words was better from Trials 1 to 5 compared to low-frequency words. Regardless of the time of recall, controls better recalled words with large neighborhood sizes, whereas persons with DAT showed the opposite pattern. Word length had no effect on learning and recall. The implications pertain to three areas: (1) psycholinguistic differences between controls and person with DAT, (2) potential clinical and diagnostic applications of these differences, and (3) the construction and evaluation of verbal memory tests.

Our study was conducted with the word-list A of the first parallel form of the CVLT-G, which is used in Germany, Austria, and Switzerland. However, it is unlikely that our findings only pertain to the German language. Linguistic recall effects are assumed to reflect cognitive processes that underlie verbal memory (Roodenrys et al., 2002), which should be universal. The frequency effect, for example, seems to be active at basic physiological levels of processing. During reading, the fixation time for high-frequency words is shorter than for low-frequency words, indicating faster processing of the former (Inhoff & Rayner, 1986) and low-frequency words require more effort than high-frequency words during encoding (Diana & Reder, 2006). Also, the importance of word frequency for a range of verbal and lexical tasks has been demonstrated across several languages (Brysbaert et al., 2011) and our findings concur with international experimental studies (Hicks, Marsh & Cook, 2005; Roodenrys et al., 2002). Therefore, it seems more plausible that our results pertain to basic cognitive processes instead to only the German language. To confirm these assumptions, future studies are needed to replicate the findings in different languages. English databases are available online with CELEX (celex.mpi.nl) and MCWord (neuro.mcw.edu/mcword). Overviews of linguistic databases and their specificities are provided by the Brigham Young University (corpus.byu.edu/overview.asp) and the University of Wollongong (uow.edu.au/~dlee/CBLLinks.htm).

The fact that patients were matched for depressive symptoms and underwent state-of-the-art diagnostics strongly reduces the chance of mistaking depression for dementia. It is, therefore, unlikely that the results of this study are distorted by the patients' mood.

Differential Linguistic Recall Effects for Controls and Persons with DAT

The observed recall advantage for high-frequency words is in line with some experimental studies that found a recall advantage for high-frequency words (Roodenrys et al., 2002; Hicks, Marsh & Cook, 2005) yet, contradicts studies that found the opposite (DeLosh & McDaniel, 1996; Ozubko & Joordens, 2007). When words are presented for recall, they are thought to leave a verbal memory trace that decays over time but can be strengthened by the words linguistic characteristics. During recall, this trace is picked up and the full word can be reconstructed by the process of redintegration (Schweickert, 1993). Redintegration describes the reconstruction of the whole from a part of it. In this study, the whole refers to the complete word whereas the part alludes to its decaying memory trace. High-frequency words, for example, are thought to have a recall advantage as they leave a stronger memory trace and better facilitate redintegration compared to low-frequency words (Roodenrys et al., 2002). There were no effects of diagnosis on an advantage for either low or high-frequency words in recognition. Numerically, however, persons without DAT showed an advantage for high-frequency words, whereas there was no difference in recognition rates between low- and high-frequency words in persons with DAT.

The results of our study suggest that the frequency-based redintegration may still function in DAT. The difference in recall rates between high- and low-frequency words was smaller in persons with DAT, but still reached statistical significance. The decrease in the difference from controls to persons with DAT, however, may point to a breakdown in redintegration as a consequence of Alzheimer's disease. As indicated by the results of the CERAD test battery and considering that the persons with DAT were diagnosed for the first time, it can be assumed that the Alzheimer disease was at mild or moderate stages. It remains to be investigated whether there is no difference in recall rates of high- and low-frequency words at more advanced stages of DAT, which would indicate a complete disruption of redintegration.

Strikingly, controls and persons with DAT showed directly opposite recall patterns with regard to neighborhood size across all times of recall and, though not statistically significant, in their recognition performance. Whereas the controls showed the expected effect of better recall for words with large neighborhood sizes that has also been found in experimental studies (Jalbert et al., 2011b; Roodenrys et al., 2002), the persons with DAT better recalled words with small neighborhood sizes. It has been argued that large neighborhood sizes may facilitate encoding and recall by activation of a larger associative network of neighbors (Jalbert, Neath & Surprenant, 2011a). Similar to the word frequency effect, this process has been explained by redintegration with large neighborhood words leaving better traces for later retrieval (Roodenrys et al., 2002). As persons with DAT showed the reverse pattern and the difference in recall rates between small and large neighborhood words was statistically significant in this group, disrupted redintegration in DAT may not alone underlie the interaction between diagnosis and neighborhood size. Rather, it is possible that a decrease in neural connectivity of verbal memory systems underlies these observed differences (Wolk, Dickerson & the Alzheimer's Disease Neuroimaging Initiative, 2011). Whereas the breakdown of associative language networks could account for the absence of an advantage for words with large neighborhoods, it cannot explain why the recall rates dropped below those of small neighborhood words. These words might bear some features that support a memory process that is relied on in DAT.

An effect that is similar with regard to the mirrored pattern has been found in the recognition of low-frequency words. Even though this was not confirmed in this study, it has been suggested that low-frequency words are usually better recognized by healthy controls, whereas persons with DAT have a markedly lower hit rate for these words with the rates for high-frequency words remaining similar to those of healthy controls (Balota et al., 2002; Wilson et al., 1983). Potentially, the two groups relied on different memory processes to make a judgment. Whereas the controls may have used recollection (explicit recall of the stimulus) to decide about ruling a word in or out, the persons with DAT might have built on familiarity (the feeling of having encountered the stimulus before) (Yonelinas, 2002). The controls, therefore, recognized the seldom encountered words, which were particularly salient to them. The persons with DAT, in contrast, were not able to use recollection and employed familiarity, which leads to a better recognition of often heard and read words. It is possible that a comparable dissociation of neighborhood size and verbal memory occurs in persons with DAT, rendering them unable to draw on commonly used memory processes. Whereas differences in recognition memory have already been mapped, experimental evidence is needed to investigate the qualitative linguistic changes in recall that occur in DAT. An investigation of further CVLT-scores like the number of intrusions and their association with linguistic factors might help to differentiate underlying memory processes that might be differently affected in controls and persons with DAT.

Theoretical Diagnostic Implications

There is little evidence on the interaction between mental status and linguistic verbal memory effects. Whereas no differential associations of word frequency with recall and recognition were found in a study comparing persons with and without schizophrenia (Brébion, David, Bressan & Pilowsky, 2005), alcohol and lorazepam reduced the advantage for high- over low-frequency words in immediate and free recall in mixed lists compared to a placebo control group (Soo-ampon, Wongwitdecha, Plasen, Hindmarch & Boyle, 2004). Interestingly, this study also found that recall rates decreased stronger for low-frequency words than for high-frequency words with increasing levels of sedation. This suggests that recall of low-frequency words, which already leave a weaker memory trace, might be earlier compromised by mental changes and, therefore, bear potentially important diagnostic information. In this study, persons with DAT showed a recall advantage for high-frequency words; however, the effect was markedly smaller than for the controls. A decreased or absent word frequency effect might indicate cognitive impairment.

If future investigations confirmed the different recall patterns for controls and persons with DAT according to neighborhood size, this effect might be of diagnostic use. Whereas a positive difference between the recall rates of large minus small neighborhood words might indicate normal verbal memory functioning, a negative difference might be a sign of impairment. Linguistic parameters might consequently be used to increase a tests diagnostic accuracy. However, much more research is needed to investigate the nature of these effects before any clinical application is justified.

Implications for Verbal Memory Test Construction and Interpretation

When devising a new verbal memory test, the possibilities for real innovation are limited. Verbal memory functioning can only be assessed by visually or orally presenting verbal information and subsequently testing oral or written recall performance and auditory or visual word recognition. To distinguish novel from existing tests, several variables can be altered, including the number and type of subtests, as well as interval and list length. Even though common practice in the construction of other cognitive tests, the difficulty of the individual items has seldom been a factor in the construction of verbal memory tests. Choosing words according to their linguistic characteristics offers a novel way of controlling the difficulty and diagnostic fairness of word-list learning tests. Two examples based on the results of this study are given. (1) As they appear to have lower recall rates, only low-frequency words could be selected for a rather challenging test that aims to avoid ceiling effects and discriminate well in higher ranges of performance. The opposite was done with the CVLT-II to create easier lists. (2) As there seems to be a difference between controls and persons with DAT, neighborhood size could be kept as stable and homogenous as possible across words in order to prevent favoring one group by an overrepresentation of, for example, large neighborhood words.

Another factor of relevance for the clinical setting and test construction is a possible interaction of linguistic effects with serial position, even though this did not pertain to the learning list A of the CVLT-G. There is evidence that frequency effects on recall may particularly occur at recency positions (Van Overschelde, 2002). An uneven distribution of word characteristics across serial positions could challenge the diagnostic meaningfulness of serial cluster scores as well as primacy and recency effects, as recall advantages of certain positions might be confounded with underlying word characteristics. Particular consequences might arise for the assessment of verbal memory functioning in late life cognitive decline. In healthy subjects, word-list recall usually follows a U-shape, indicating that words from primacy and recency positions are better recalled than words from intermediate positions (Murdock, 1962). An important and early neuropsychological marker in mild cognitive impairment and dementia is a weakened primacy effect and a more pronounced recency effect (Egli et al., 2014). Because persons with depression usually display the U-shaped recall pattern, serial position scores are useful in the differential diagnostics of depression and dementia (Foldi et al., 2003). An uneven distribution of linguistic characteristics across serial positions might attenuate or mask these diagnostically valuable differences.

Strengths and Limitations

To the best of our knowledge, this study was the first to examine psycholinguistic recall effects with an established instrument for the assessment of verbal memory in a clinical setting. This actuality accounts for both the strengths and weaknesses of our study. The investigation in an applied setting increases the study's ecological validity and relevance for the clinical use of verbal memory tests. The generalizability of the results to non-clinical settings might be limited by the fact that the participants were outpatients of a memory clinic with and without diagnoses of DAT. Even though these patients underwent a thorough state-of-the-art diagnostic work-up, it cannot be excluded that some of the controls might have been in prodromal phases of cognitive decline and group differences might be smaller than in experimental studies. Because we wanted to examine the CVLT-G in the clinical setting, however, this was inherent to the study's design. The classification of the words with regard to frequency and neighborhood size depends on size and compilation of the employed text corpus of the linguistic database. The dlexDB, which was employed for the linguistic analysis, is based on a diverse and recent corpus that has the largest volume of the German databases. Also, the findings may only apply to the list A of the first parallel version of the CVLT-G.

Conclusion

This study found evidence that the German CVLT might be subject to linguistic recall effects that in part present differently for persons with DAT and across time. These results are in line with international experimental studies and are likely not specific to the German language.

Conflict of Interest

None declared.

Acknowledgements

We are grateful to past and present staff of the memory clinic and the neuropsychological unit at the Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technical University of Munich, for conducting the neuropsychological and medical examinations that form the basis of this study.

References

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: Evidence for two processes in episodic recognition performance. *Journal of Memory and Language*, *46*, 199–226.

Bäckman, L., Jones, S., Berger, A.-K., Laukka, E. J., & Small, B. J. (2005). Cognitive impairment in preclinical Alzheimer's disease: A meta-analysis. *Neuropsychology*, *19*, 520–531.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonia, TX: The Psychological Corporation.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289–300.

Bireta, T. J., Neath, I., & Surprenant, A. M. (2006). The syllable-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review*, *13*, 434–438.

Bondi, M. W., Jak, A. J., Delano-Wood, L., Jacobson, M. W., Delis, D. C., & Salmon, D. P. (2008). Neuropsychological contributions to the early identification of Alzheimer's disease. *Neuropsychology Review*, *18*, 73–90.

Brébion, G., David, A. S., Bressan, R. A., & Pilowsky, L. S. (2005). Word frequency effects on free recall and recognition in patients with schizophrenia. *Journal of Psychiatric Research*, *39*, 215–222.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental Psychology*, *58*, 412–424.

Criss, A. H., Aue, W. R., & Smith, L. (2011). The effect of word frequency and context variability in cued recall. *Journal of Memory and Language*, *64*, 119–132.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *Manual for the California Verbal Learning Test - Second Edition (CVLT–II)*. San Antonio, TX: The Psychological Corporation.

DeLosh, E. L., & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1136–1146.

Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 805–815.

Diehl, J., & Kurz, A. (2002). Frontotemporal dementia: Patient characteristics, cognition, and behaviour. *International Journal of Geriatric Psychiatry*, *17*, 914–918.

Egli, S. C., Beck, I. R., Berres, M., Foldi, N. S., Monsch, A. U., & Sollberger, M. (2014). Serial position effects are sensitive predictors of conversion from MCI to Alzheimer's disease dementia. *Alzheimer's & Dementia*, *10* (S), S420–S424.

Foldi, N. S., Brickman, A. M., Schaefer, L. A., & Knutelska, M. E. (2003). Distinct serial position profiles and neuropsychological measures differentiate late life depression from normal aging and Alzheimer's disease. *Psychiatry Research*, *120*, 71–84.

Gainotti, G., Quaranta, D., Vita, G., & Marra, C. (2014). Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease. *Journal of Alzheimer's Disease*, *38*, 481–495.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*, 23–41.

Glanc, G. A., & Greene, R. L. (2007). Orthographic neighborhood size effects in recognition memory. *Memory & Cognition*, *35*, 365–371.

Guo, Y., Logan, H. L., Glueck, D. H., & Muller, K. E. (2013). Selecting a sample size for studies with repeated measures. *BMC Medical Research Methodology*, *13*, 100.

Hautzinger, M., Bailer, M., Worall, H., & Keller, F. (1995). *Beck-Depressions-Inventar (BDI) Manual*. (2nd ed.). Bern: The Huber.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, et al. (2011). dlexDB – A lexical database for the psychological and linguistic research (article in German). *Psychologische Rundschau*, *62*, 10–20.

Hicks, J. L., Marsh, R. L., & Cook, G. I. (2005). An observation on the role of context variability in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1160–1164.

Hulme, C., Suprenant, A. M., Bireta, T. J., Stuart, G., & Neath, I. (2004). Abolishing the word-length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 98–106.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, *40*, 431–439.

Jahn, T., Theml, T., Diehl, J., & Grimmer, T. (2004). CERAD-NP and flexible battery approach in the neuropsychological differential diagnosis of dementia versus depression (article in German). *Zeitschrift für Gerontopsychologie & -psychiatrie*, *17*, 77–95.

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011b). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 338–353.

Jalbert, A., Neath, I., & Surprenant, A. M. (2011a). Does length or neighborhood size cause the word length effect? *Memory & Cognition*, *39*, 1198–1210.

Katkov, M., Romani, S., & Tsodyks, M. (2014). Word length effect in free recall of randomly assembled word lists. *Frontiers in Computational Neuroscience*, *8*, 1–4.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, *10*, 707–710.

Lohnas, L. J., & Kahana, M. J. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1943–1946.

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22* (1), 132–142.

Mannhaupt, H.-R. (1983). German category norms for verbal items in 40 categories (article in German). *Sprache & Kognition*, *2*, 264–278.

Morris, J. C., Heyman, A., Mohs, R. C., & Hughes, J. P. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, *39*, 1159–1165.

Murdock, B. B. Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Niemann, H., Sturm, W., Thöne-Otto, A., & Willmes, K. (2008). *California Verbal Learning Test. Deutsche Adaptation*. Frankfurt am Main: Pearson Assessment & Information GmbH.

Ozubko, J. D., & Joordens, S. (2007). The mixed truth about frequency effects on free recall: Effects of study list composition. *Psychonomic Bulletin & Review*, *14*, 871–876.

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1019–1034.

Rubin, D. C., & Friendly, M. (1986). Predicting which words get recalled: Measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns. *Memory & Cognition*, *14*, 79–94.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, *21*, 168–175.

Soo-ampon, S., Wongwitdecha, N., Plasen, S., Hindmarch, I., & Boyle, J. (2004). Effects of word frequency on recall memory following lorazepam, alcohol, and lorazepam alcohol interaction in healthy volunteers. *Psychopharmacology*, *176*, 420–425.

Thalmann, B., & Monsch, A. U. (1997). *CERAD. The consortium to establish a registry for Alzheimer's disease. Neuropsychologische Testbatterie*. Basel: Memory Clinic Basel.

Van Overschelde, J. P. (2002). The influence of word frequency on recency effects in directed free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 611–615.

Wilson, R. S., Bacon, L. D., Fox, J. H., & Kramer, R. L. (1983). Word frequency effect and recognition memory in dementia of the Alzheimer type. *Journal of Clinical and Experimental Neuropsychology*, *5*, 97–104.

Wolk, D. A., Dickerson, B. C., & the Alzheimer's Disease Neuroimaging Initiative (2011). Fractionating verbal episodic memory in Alzheimer's disease. *Neuroimage*, *54*, 1530–1539.

Wright, S. L., & Persad, C. (2007). Distinguishing between depression and dementia in older persons: Neuropsychological and neuropathological correlates. *Journal of Geriatric Psychiatry and Neurology*, *20*, 189–198.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.

Use of previously published paper in disseration

Lee, Gregory <GLEE@augusta.edu>

Di 03.01.2017 16:35

[03] DISS/AWLT

An: Heßler, Johannes <johannes.hessler@tum.de>;

Cc: acn-feedback@highwire.stanford.edu <acn-feedback@highwire.stanford.edu>; 'anna.hernandez-french@oup.com' <anna.hernandez-french@oup.com>;

Dear Johannes Hessler,

Oxford University Press grants permission for you to include your previously published paper in Archives of Clinical Neuropsychology within your dissertation as long as you cite that it was previously published in our journal.

Thank you,

Gregory Lee

Gregory P. Lee, PhD, ABPP
Board Certified in Clinical Neuropsychology
Editor-in-Chief, *Archives of Clinical Neuropsychology*
Professor of Neurology
Medical College of Georgia
Augusta, GA 30912
(706) 721-3851

NAME: Johannes Hessler
EMAIL: johannes.hessler@tum.de
IP ADDRESSES: 92.225.225.196, 92.225.225.196
HOSTNAME: x5ce1e1c4.dyn.telefonica.de
PREVIOUS PAGE: http://acn.oxfordjournals.org/
BROWSER: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/53.0.2785.143 Safari/537.36, oupjournals-cluster
PROMOTIONAL USE: Granted
SESSION ID: ViNUmJnKJT-To9LZNKv46A

COMMENTS:
Dear madam or sir,

my paper "differential linguistic recall effects in the California Verbal Learning Test..." was published in the Archives of Clinical Neuropsychology. I am currently working on my publication-based dissertation, in which I want to include this paper. For this, I am required to get permission from the journal/publisher, as the dissertation will later be uploaded on my university's library website. Therefore, I wanted to inquire the general policy of ACN/Oxford Journals and, if possible, request a statement that allows me to imbed my publication into my dissertation.

Thank you for your efforts, kind regards, and merry christmas
Johannes Hessler
Oxford University Press (UK) Disclaimer

Appendix C

Applying psycholinguistic evidence to the construction of a new test of verbal memory in late-life cognitive decline: the Auditory Wordlist Learning Test (AWLT)

Assessment
1–13
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191117690603
journals.sagepub.com/home/asm
SAGE

Applying Psycholinguistic Evidence to the Construction of a New Test of Verbal Memory in Late-Life Cognitive Decline: The Auditory Wordlist Learning Test

Johannes Baltasar Hessler[1], David Brieber[2], Johanna Egle[2], Georg Mandler[2], and Thomas Jahn[1]

Abstract

The construction of the German Auditory Wordlist Learning Test (AWLT) for the assessment of verbal memory in late-life cognitive decline was guided by psycholinguistic evidence, which indicates that a word's linguistic characteristics influence its probability of being learned and recalled. The AWLT includes four trials of learning, short and long delayed free recall, and a recognition task. Its words were selected with taking into account their semantic content, orthographic length, frequency in the language, and orthographic neighborhood size (the number of words derived by adding, subtracting, or replacing a single letter at a time). Through this method, it was possible to better control item and test difficulty, improve the similarity between parallel forms, and reduce bias through recall advantages for certain words due to their linguistic characteristics. In two pilot studies with cognitively healthy subjects, the AWLT showed good internal consistency, split-half reliability, and parallel forms reliability and proved able to assess learning, retention, and recognition. Overall, linguistic recall effects were mitigated; however, an advantage for high-frequency words was observed.

Keywords

verbal memory, wordlist, linguistics, word length, frequency, neighborhood size

Item difficulty is usually a central concern in the construction of cognitive tests. In verbal memory tests using the wordlist paradigm, however, this issue is often neglected or hardly considered. This is surprising given that there is evidence indicating that, among others, variables like a word's length, frequency in the language, and neighborhood size can influence its probability of being learned, recalled, and recognized.

Word length can be operationalized as the number of a word's syllables (Jalbert, Neath, Bireta, & Surprenant, 2011) and usually short words are better recalled than long words (e.g., Hulme, Suprenant, Bireta, Stuart, & Neath, 2004; Jalbert et al., 2011). Word frequency indicates how common a word is in a given language (Criss, Aue, & Smith, 2011). High-frequency words are often better recalled, while low-frequency words are better recognized (e.g., MacLeod & Kampe, 1996). Neighborhood size describes the number of words that can be created by replacing, adding, or deleting a single letter of a word at a time. Large neighborhood sizes are usually related to a higher recall probability (e.g., Jalbert et al., 2011; Roodenrys, Hulme, Lethbridge, Hinton, & Nimmo, 2002).

These findings have three important implications for the construction of verbal memory tests using the wordlist learning paradigm. First, linguistic characteristics could be employed to control the difficulty of such tests on the item level. This was done in the construction of the second version of the California Verbal Learning Test (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000), for which only high-frequency words were selected to render the list easier to learn and recall. Second, matching wordlists with regard to their linguistic profiles allows for increasing the similarity between parallel forms and between learning and distraction lists in the recognition task. Third, there is evidence indicating an interaction between the linguistic profile of wordlists and the cognitive status of the test takers. For example, the recall advantage for high- over low-frequency words was reduced in subjects sedated with lorazepam and alcohol compared with controls (Soo-ampon, Wongwitdecha, Plasen, Hindmarch, &

[1]Technical University of Munich, Munich, Germany
[2]Schuhfried GmbH, Mödling, Austria

Corresponding Author:
Johannes Baltasar Hessler, Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technical University of Munich, Ismaninger Strasse 22, 81675 Munich, Germany.
Email: johannes.hessler@tum.de

Boyle, 2004). Furthermore, the German version of the CVLT (Niemann, Sturm, Thöne-Otto, & Willmes, 2008) was found to be subject to linguistic recall effects (Hessler, Fischer, & Jahn, 2016). While both controls and persons with Alzheimer's dementia recalled more high- than low-frequency words, there was a mirrored pattern with regard to neighborhood size. The controls had significantly higher recall rates for words with large neighborhood sizes, while the opposite was true for persons with dementia. These interactions might bias the diagnostic accuracy of a wordlist learning test. A list containing many words with small neighborhood sizes, for example, could support the memory performance of persons with dementia and, thereby, obscure group differences and lead to reduced diagnostic accuracy of the test. A possible way to mitigate these effects would be to reduce the variability in the linguistic factors between the words in order to create a list that comprises highly similar words.

Controlling the linguistic properties of the selected words also allows for increasing equivalence in the translation of existing tests. While, for example, there is only little information about how the wordlist within the neuropsychological test battery of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD; Morris, Heyman, Mohs, & Hughes, 1989; CERAD-WL) was compiled, its Korean translation is linguistically closely matched (Lee et al., 2002). The Korean CERAD-WL resembles the original version with regard to the relative word frequency, semantic category, and partly in word length. For the German version, the original words were directly translated (Thalmann & Monsch, 1997).

The present study describes the construction of the German Auditory Wordlist Learning Test (AWLT; German: *Auditiver Wortlisten Lerntest*) that was based on psycholinguistic evidence. Furthermore, the AWLT's psychometric qualities and the presence of linguistic recall effects were investigated in two pilot studies with cognitively healthy subjects. The AWLT was developed as a cooperative project by the Schuhfried GmbH (Mödling, Austria) and the Clinical and Experimental Neuropsychology unit of the Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technical University of Munich (Germany). The AWLT will be part of the tablet-based neuropsychological test battery Cognitive Functions Dementia (CFD) for the diagnosis of dementia that is currently being normed and validated by the developers. The battery is the first tablet-based test set within the well-known Vienna Test System and aims at the early identification and differential diagnosis of predominantly neurodegenerative dementia syndromes.

Construction of the AWLT

General Aims of Test Construction

The above described psycholinguistic knowledge was applied to the construction of the AWLT to (a) control the test's difficulty on the item level (linguistic item difficulty), (b) increase the similarity between parallel forms as well as between learning and distraction lists, and (c) mitigate linguistic recall effects as much as possible. Furthermore, the AWLT was designed to be a valuable alternative to existing wordlist-learning tests in terms of its structure and linguistic profile.

Commonly employed tests of verbal memory in the context of aging, mild cognitive impairment, and dementia are the CVLT and the CERAD-WL. While the CVLT might be too exhausting for some patients, the CERAD-WL might be too easy in certain cases and produce ceiling effects. To construct a valuable alternative for these two established tests, we aimed to place the AWLT between CVLT and CERAD-WL with regard to its length and number of measures. Assuming that longer paradigms with more measures are more demanding for test takers, we aimed to develop the AWLT as an intermediate solution. Of course, the AWLT might as well be employed in the diagnostics of conditions other than dementia.

Structure of the AWLT

The AWLT has two parallel forms, each comprising 12 learning words. The AWLT's structure consists of four measures (see Table 1 for a comparison with CERAD-WL and CVLT).

1. *Learning Phase*: During the four trials of the learning phase, the 12 words are read to the test taker, who is to recall as many words as possible after each trial with the original order being irrelevant.
2. *Short Delayed Free Recall:* After an interval of 5 minutes filled with nonverbal and nonmemory-related tests and without being previously warned, the test taker is again to recall as many words as possible with the original order being irrelevant.
3. *Long Delayed Free Recall:* After an interval of 20 minutes filled with nonverbal as well as nonmemory-related tests and without being previously warned, the test taker is again asked to recall as many words as possible with the original order being irrelevant.
4. *Recognition:* A list of 24 words containing the original 12 and 12 new but semantically matched words is read to the test taker, who has to recognize the learned words among the distractors.

The AWLT ranges between the CVLT and the CERAD-WL with regard to number of words and number of measures (Table 1). Similar to the CVLT, the words are read to the subject. As with the CERAD-WL, the AWLT does not include a second learning list.

The CERAD-WL requires that the words are read aloud by the test taker, which ensures that the words are perceived as well as encoded and prevents the use of rehearsal strategies.

Table 1. Test Structures of CERAD-WL, AWLT, and CVLT.

Parameter	CERAD-WL	AWLT	CVLT
Number of parallel forms	1	2	3
Number of words	10	12	16
Presentation of words	Visual and read aloud by patient	Read to patient	Read to patient
Number of trials in learning phase	3	4	5
Distracting list	No	No	Yes
Short delayed free recall	No	Yes	Yes
Short delayed cued recall	No	No	Yes
Long delayed free recall	Yes	Yes	Yes
Long delayed cued recall	No	No	Yes
Recognition	Yes	Yes	Yes
Forced-choice recognition	No	No	Yes

Note. CERAD-WL = Consortium to Establish a Registry for Alzheimer's Disease neuropsychological test battery wordlist; CVLT = California Verbal Learning Test; AWLT = Auditory Wordlist Learning Test.

As a consequence, recall scores are assumed to reflect "real" recall performance that is unaffected by the use of strategies. As the AWLT was intended as a purely auditory test, we decided to have the words read to the subject by the examiner or played from an audio file on the tablet. We think that, for two reasons, auditory word presentation does not introduce more difficulties with regard to rehearsal and encoding than visual presentation: (a) Visual presentation cannot completely prevent the use of strategies. For example, it is possible to conceive a story, which develops along the wordlist as it is read and that can later be reconstructed to promote recall. (b) Encoding can also be assessed and distinguished from retrieval with a test employing auditory word presentation. Impaired encoding can be suspected when a specific profile of low performance in learning, recall, and recognition is observed (Delis et al., 1991; Miller, 1956). In the learning phase, encoding deficits are signified by no or very little improvement over the individual trials and recall rates that do not exceed the auditory working memory of 7 ± 2 items. When these deficits co-occur with low performance at delayed recall and recognition trials, the inability to encode the words and move them to long-term memory can be assumed. Also, a comparison of performance between the delayed recall and recognition measures allows for discriminating impairments in retrieval and encoding. While impaired recall and intact recognition point to a retrieval deficit, an impairment of both retrieval and recognition indicates an encoding deficit (Butters, 1985).

Word Selection and List Compilation

As with the test structure, the aim for the word selection was to place the AWLT between CERAD-WL and CVLT with regard to mean word length, frequency, and neighborhood size. Furthermore, we aimed to minimize the linguistic variability of the AWLT as much as possible in order to attenuate or even prevent linguistic recall effects.

The linguistic analyses of the German CVLT's Wordlist A and the German CERAD-WL, as well as for the word selection for the AWLT, were performed with the dlexDB (Heister et al., 2011). The dlexDB is a German lexical database that is based on the text corpus of the Digital Dictionary of the German Language (*Digitales Wörterbuch der deutschen Sprache*; Geyken, 2007), which includes 122,816,010 tokens and 2,224,542 types. The sequence AABBCCDD includes 8 tokens (AABBCCDD; i.e., concrete occurrences of a word in the corpus) and 4 types (A, B, C, D; i.e., class of words). Considering a simplified example with a hypothetical corpus that includes only two sentences: (a) "Anna offers the dog a treat" and (b) "The dog eats the treat off Anna's hand." This corpus would have 14 tokens (Anna; offers; the; dog; a; treat; the; dog; eats; the; treat; off; Anna's; hand) and 10 types (a; Anna; Anna's; dog; eats; hand; off; offers; the; treat) with the tokens counting each single word in the corpus regardless whether it had occurred before or not and the types counting only the first appearance but no further ones. The Digital Dictionary of the German Language corpus comprises prose, newspaper articles, functional texts, and transcribed spoken language from the whole 20th century in equal shares. The dlexDB is free of charge and accessible online (http://www.dlexdb.de/).

Length was defined by the number of syllables. Normalized annotated type frequency, as well as normalized neighborhood size, were determined by means of the dlexDB. Normalization in the dlexDB is a form of standardization. In the case of frequency, normalized values indicate the type frequency per million tokens in the corpus. For neighborhood size, normalized values indicate the number of neighbored types per million types in the corpus. Annotation allows for analyzing orthographically similar words separately according to the different parts of speech they occupy. Through this method, we were able to extract the normalized frequency of the nouns while excluding personal names and other parts of speech. We employed normalized values for frequency and neighborhood size so that the two variables could be analyzed and interpreted on the same scale. In the remainder, "frequency" will denote annotated normalized frequency, and "neighborhood size" will denote normalized neighborhood size.

The CERAD-WL had a mean word length of 1.7 (*SD* = 0.67, range 1-3), mean frequency of 29.47 (*SD* = 27.78, range 7.19-14.22), and mean neighborhood size of 13.27 (*SD* = 7.98, range 2.57-25.69). The Learning List A of the German CVLT had a mean word length of 2.13 (*SD* = 0.89, range 1-4), mean frequency of 1.88 (*SD* = 1.83, range 0.02-5.28), and mean neighborhood size of 5.78 (*SD* = 6.17, range 0.00-8.41). As the two tests are rather opposed with regard to their mean frequency and neighborhood size, the AWLT could potentially be positioned in between. As the mean word lengths were very close to each other, we decided to select only one- or two-syllabled words, which would produce a similar mean but less variability.

The AWLT's words for the four lists were selected by a five-step procedure:

1. Predefinition of 12 semantic categories: furniture, food, transportation, clothing, tools, recreation, animals, plants, buildings, musical instruments, kitchen, and daily life.
2. Creation of a pool of words that are nouns, easily imaginable and concrete, assignable to one of the 12 semantic categories, and mono- or disyllabic.
3. Analysis of all words in the pool with regard to frequency and neighborhood size with the dlexDB.
4. Selection of 48 words that have a word frequency and neighborhood size between 5 and 15 to produce a maximal range of 10 (i.e., smaller than in CVLT and CERAD-WL) in frequency and neighborhood size.
5. Distribution of these words across four lists of 12 words (12 learning words and 12 distractors for each parallel form) in a way that the lists' mean normalized frequency and normalized neighborhood size would lie around 10 (between CVLT and CERAD-WL) and that each semantic category appears only once on each list (i.e., no semantic overlap).

It was not fully possible to select only words with frequencies and neighborhood sizes between 5 and 15. In some cases, these thresholds had to be crossed in order to fill the four lists with suitable words. The resulting lists, however, met the previously set criteria of including only one- or two-syllabled words that belong to 12 different semantic categories and having a mean frequency and neighborhood size as well as a difference between maximum and minimum around 10. The words were then randomly sorted within the lists and the lists were randomly assigned their position in the test (learning list or distraction list for the recognition trial and Forms 1 or 2).

Pilot Study 1

In the first pilot study, a paper-and-pencil version of the AWLT was administered to cognitively healthy subjects in the testing center of the Schuhfried GmbH to investigate the test's feasibility and identify potential areas for adjustment and improvement.

Subjects and Procedures

The subjects were recruited by means of newspaper advertisements in the Vienna area. All interested persons were questioned about the presence of psychiatric disorder or neurological disease and the use of neurotropic drugs. The inclusion criteria for participation were age of 16 years or older and no previous testing with a wordlist learning test within the past year. The AWLT was administered by trained staff of the testing center. The retention intervals before short and long delayed recall were filled with nonverbal and nonmemory-related neuropsychological tests.

All persons who volunteered could be included in the study. Thirty-four persons (17 female, 50.0%) were tested with the AWLT's first parallel form. Their mean age (*SD*) was 49.56 (15.86) years. Four (11.8%) finished compulsory primary education or middle school, 19 (55.9%) vocational training, 7 (20.6%) higher schools, and 4 (11.8%) university. Thirty-five persons (18 female, 51.4%) were tested with the AWLT's second parallel form. Their mean age (*SD*) was 49.74 (15.38). Two (5.7%) finished compulsory primary education or middle school, 22 (62.9%) vocational training, 6 (17.1%) higher schools, and 5 (14.3%) university.

Statistical Analysis, Results, and Discussion

The AWLT's feasibility was operationalized as the rate of tests that could be fully administered so that all test scores could be calculated. Feasibility was 100% for both forms.

Intrusions (i.e. falsely "recalled" words) were analyzed with regard to their semantic content to identify items that might produce intrusions from the same semantic category. In Form 1, 11 different intrusions were named. While "Pfeil" (arrow) was named twice, "Mühle" (mill), "Ast" (branch), "Bild" (picture), "Dampf" (steam), "Vogel" (bird), "Kugel" (ball, sphere), "Saum" (seam), "Veilchen" (violet), "Blume" (flower), and "Stuhl" (chair) occurred once. In Form 2, 11 different intrusions were named. While "Blume" (flower) occurred five times and "Kork" (cork), "Haus" (house), "Blüten" (blossoms), "Zucker" (sugar), "Puppe" (doll), "Ball" (ball), "Topf" (pot), "Dach" (roof), and "Schlüssel" (key) were named once.

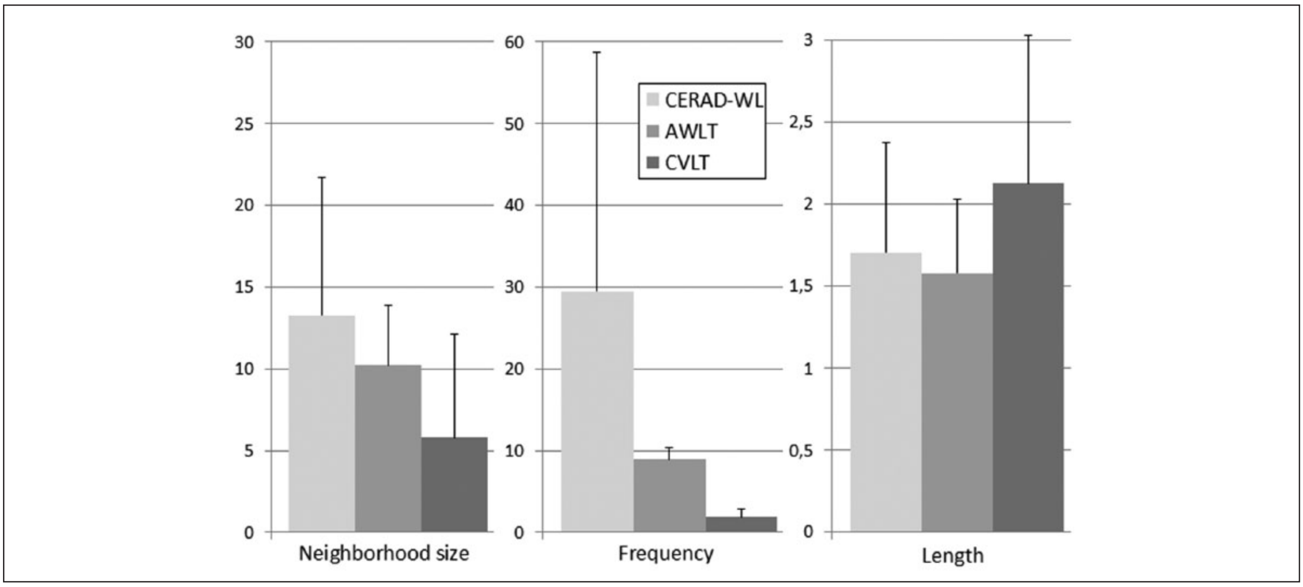


Figure 1. Means and standard deviations of length, frequency, and neighborhood size for CERAD-WL, and the first form learning lists of AWLT and CVLT.

Note. CERAD-WL = Consortium to Establish a Registry for Alzheimer's Disease neuropsychological test battery wordlist; CVLT=California Verbal Learning Test; AWLT = Auditory Wordlist Learning Test.

With five entries in 35 persons, the intrusion "Blume" (flower; frequency = 10.37, neighborhood size = 11.13) might be a result of the word "Blüte" (blossom; frequency = 14.73, neighborhood size = 6.42) that was on the learning list of the second form. Presumably, "Blume" is more prototypical and therefore a common intrusion. In order to remove this bias, "Blüte" needed to be replaced by a more suitable alternative.

Adjustment and Final Version of the AWLT

Based on the results of Pilot Study 1, the AWLT's second parallel form was adjusted. "Blüte" was replaced by "Blume," which did not affect the overall linguistic profile of the list. Form 1 remained unchanged.

The AWLT's final version was then examined with regard to linguistic item difficulty and similarity within and between parallel forms.

Statistical Analysis

Several statistical analyses were conducted to assess the AWLT's linguistic item difficulty in comparison with CVLT and CERAD-WL, as well as the similarity between its parallel forms.

Linguistic Item Difficulty. A 3 × 3 multivariate analysis of variance (MANOVA) with the between-factor *list* (AWLT Form 1; CERAD-WL; CVLT List A) and the within-factor *linguistic* (length, frequency, neighborhood size) was employed to compare the tests' linguistic profiles. Furthermore, using AWLT data from the pilot study, the recall rates for each word at Trials 1 and 4 of the learning phase, as well as at short and long delayed recall, as indicators of retention, were calculated and plotted in a line graph to investigate the presence of primacy and recency effects.

Similarity Within and Between Parallel Forms. The linguistic similarity between the four wordlists of the AWLT was tested with a 4 × 3 MANOVA, with the between-factor *list* (Form 1 learning; Form 1 distraction; Form 2 learning; Form 2 distraction) and the within-factor *linguistic* (length; frequency; neighborhood size). It is not possible to confirm a null hypothesis by means of statistical hypothesis testing. A failure to reject the null hypotheses of equal mean values would only suggest a high probability that the lists are linguistically not different. Therefore, interpretations were mainly based on effect sizes of the differences.

Significant main effects of MANOVAs were decomposed with *t* tests. For post hoc tests, 95% confidence intervals for the mean differences and Cohen's *d* as a measure of effect size are given. *p* Values were adjusted with the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995). Data analysis was performed with SPSS 23.

Results

Linguistic Item Difficulty and Comparison With CVLT and CERAD-WL. Figure 1 displays mean length, frequency, and neighborhood size for the CERAD-WL and the learning

Table 2. Linguistic Profile of the AWLT With Regard to Word Length, Frequency, and Neighborhood Size.

	Form 1		Form 2	
Linguistic characteristic	Learning	Distraction	Learning	Distraction
Word length				
M (*SD*)	1.75 (0.13)	1.75 (0.13)	1.58 (0.15)	1.58 (0.15)
Minimum	1	1	1	1
Maximum	2	2	2	2
Range	1	1	1	1
Frequency				
M (*SD*)	9.59 (0.94)	8.34 (1.18)	8.91 (0.71)	8.70 (0.93)
Minimum	4.83	2.67	5.50	4.14
Maximum	15.10	14.22	13.95	14.83
Range	10.27	11.55	8.45	10.69
Neighborhood size				
M (*SD*)	10.31 (1.06)	9.34 (0.94)	10.20 (1.02)	8.74 (0.89)
Minimum	5.57	3.85	5.99	5.14
Maximum	16.27	13.70	15.41	14.98
Range	10.70	9.85	9.42	9.84

Note. AWLT = Auditory Wordlist Learning Test; *M* = mean; *SD* = standard deviation.

lists of the first forms of AWLT and CVLT. While the word lengths were close to each other, the AWLT's mean frequency and neighborhood size lay between the other two tests. The dispersion of the linguistic variables was the smallest for the AWLT, except for frequency, which showed similarly little spread in the CVLT.

MANOVA comparing length, frequency, and neighborhood size of the three tests revealed a difference in the combination of the three variables, Wilks's $\Lambda = 0.55$, $F(6, 66) = 3.80$, $p = .003$, $\eta^2 = 0.26$. Tests of between-subject effects revealed significant differences in frequency, $F(2, 35) = 10.54$, $p < .001$, $\eta^2 = 0.38$, and neighborhood size, $F(2, 35) = 4.62$, $p = .017$, $\eta^2 = 0.21$, but not in length, $F(2, 35) = 0.74$, $p = .253$, $\eta^2 = .08$. Post hoc independent t tests indicated that the AWLT's words had a higher frequency ($M = 9.59$, $SD = 3.27$) than the CVLT's ($M = 1.88$, $SD = 1.89$), $t(26) = 7.86$, $p < .001$, Cohen's $d = 3.01$, and had a higher neighborhood size ($M = 10.31$, $SD = 3.66$) compared with the CVLT's ($M = 5.78$, $SD = 6.37$), $t(26) = 2.20$, $p = .037$, Cohen's $d = 0.84$. The tests did not differ with regard to word length, $t(26) = -1.34$, $p =.192$, Cohen's $d = 0.52$. Even though the difference was only marginally significant, the effect size indicated that the AWLT's words had a lower frequency ($M = 9.59$, $SD = 3.27$) than the CERAD-WL's ($M = 29.47$, $SD = 29.29$), $t(9.19) = -2.14$, $p = .061$, Cohen's $d = 1.00$. The degrees of freedom of the t statistic were adjusted since the variances of frequency were not equal for AWLT and CERAD-WL. The words of the AWLT and the CERAD-WL did not differ with regard to length, $t(20) = 0.21$, $p = .838$, Cohen's $d = 0.09$, and neighborhood size, $t(11.82) = -1.10$, $p = .283$, Cohen's $d = 0.47$.

Similarity Within and Between Parallel Forms. Learning and distraction lists had the same mean word length for both parallel forms (Table 2). All lists had mean frequencies below 10 and ranges around 10. Two lists had mean neighborhood sizes below 10 and ranges were around 10. Importantly, the aim of linguistic homogeneity within and between the lists was achieved, which increases the similarity between lists and parallel forms. MANOVA revealed no differences between the lists with regard to mean length, frequency, and neighborhood size, Wilks's $\Lambda = 0.91$, $F(9, 102.37) = 0.46$, $p = .901$, $\eta^2 = 0.03$.

Discussion

The AWLT was constructed with its linguistic properties in mind. The individual wordlists are linguistically similar within and between the two parallel forms, which extends their equivalence to the item level. The AWLT's mean word frequency and neighborhood size lie between the values of CVLT and CERAD-WL, suggesting an intermediate linguistic item difficulty for the AWLT. Also, the test's length as well as its number of measures and, thereby, its demand on the test taker lies between the two established alternatives CVLT and CERAD-WL. A subsequent pilot study was conducted with the AWLT's final version.

Pilot Study 2

A second pilot study of the AWLT was conducted to investigate the AWLT's psychometric qualities, the trajectories of test scores within and between the parallel forms, and the presence of linguistic recall effects. These analyses were based on the test results obtained at the two pilot studies.

Method

Subjects and Procedures. The participants of Pilot Study 2 were a subset of the sample of Pilot Study 1. Subjects who completed the AWLT's Form 1 at Study 1 were administered the adjusted and final Form 2 at Study 2 several weeks later and those who completed the original Form 2 at Study 1 were administered Form 1 at Study 2. Otherwise the procedures were similar for the two studies.

In the second session, Form 1 was administered to 22 subjects who completed Form 2 in the first pilot. Form 2 was administered to 21 subjects who completed Form 1 in the first session. Due to the adjustments after the first session, only data from the second session will be analyzed for Form 2. As the parallel forms comprise distinct learning lists, no practice effects would be expected so that Form 1 will be analyzed with the data of both session combined. Comparisons between the parallel forms will be conducted with data from the 21 subjects who completed Form 1 and the final Form 2.

Statistical Analysis

Psychometric qualities. A range of test scores was calculated: (a) the sum of correctly recalled words at each trial of the learning phase, as well as short and long delayed recall; (b) the sum score across all trials of the learning phase; and (c) the number of true positives, false positives, true negatives, and false negatives in the recognition measure, as well as an indicator of accuracy [(true positives + true negatives)/24].

Internal consistency was assessed by Cronbach's alpha and an odd–even split-half method. Correlations between sum scores of odd and even items were calculated for all individual trials of the learning phase, the sum of the scores at the individual learning trials (1 + 3 vs. 2 + 4), and short and long delayed recall and corrected by the Spearman–Brown formula for reduced test length. Parallel-forms reliability was examined by correlating the above described test scores as well as the recognition accuracy that were obtained in the two forms.

Test scores within and between parallel forms. Mean scores on the measures of the two forms were analyzed and compared with a 7 × 2 repeated-measures analysis of variance (RM-ANOVA) with the within-factors *time* (Trial 1; Trial 2; Trial 3; Trial 4; short delayed recall; long delayed recall; true positives in recognition) and *form* (Form 1; Form 2).

Linguistic recall effects. Frequency and neighborhood size of the words of the AWLT's two parallel forms were correlated with their recall (Learning Trials 1 and 4 as well as short and long delayed free recall) and recognition rates using Spearman's rho.

The influence of length, word frequency, and neighborhood size on recall and recognition performance was investigated according to a previously employed method (Hessler et al., 2016). For that purpose, the learning list of the AWLT's first form was dichotomized to form groups of words that are relatively low or high with regard to a certain linguistic characteristic. This was done separately for length, frequency, and neighborhood size by using the median of each variable as cutoff. Recall rates for words above and below the median were then compared at Trials 1 and 4 of the learning phase, and at short and long delayed free recall. For each linguistic variable, a 4 × 2 RM-ANOVA with the two within-factors *time* (Trial 1; Trial 4; short delayed free recall; long delayed free recall) and *linguistic* (below median; above median) and the interaction between *time* and *linguistic* was conducted. Each of the RM-ANOVAs analyzes the same variance for *time*. Therefore, we corrected the p values for the main effect of time with the Bonferroni method.

A failure to reject the null hypothesis of equal recall rates was desired, as it would indicate a high probability that the linguistic characteristics do not influence recall performance. A significant time effect, however, would reflect the AWLT's ability to assess learning and retention.

Significant interaction effects in RM-ANOVAs involving linguistic variables were decomposed with t tests. For post hoc tests, 95% confidence intervals for the mean differences and Cohen's d as measure of effect size are given. p Values were adjusted with the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995). Data analysis was performed with SPSS 23.

Results

Form 1 was administered to 34 subjects (17 female, 50%) with a mean age of 49.03 (SD = 15.86) and Form 2 was administered to 35 subjects (18 female, 51.4%) with a mean age of 49.74 (SD = 15.38). Table 3 displays the characteristics of the subjects who completed the final version of the AWLT.

Psychometric Qualities. Split-half reliabilities, internal consistencies, parallel-forms reliability, and parametric correlations between the subscores are shown in Table 4. The split-half reliability and internal consistency were low at Trial 1 but increased to Trial 4 to acceptable values. The learning sum, a core variable of the AWLT, was highly reliable. At short and long delayed recall the values were also acceptable. The parallel-forms reliability was good, except at Trial 1 and for the recognition accuracy.

Table 3. Characteristics of the Participants Who Completed the Final Version of the AWLT.

Characteristic	Form 1 (*N* = 56)[a]	Forms 1 and 2 (*N* = 21)
Age; *M* (*SD*)	48.76 (15.23)	47.02 (15.13)
Female; *n* (%)	27 (48.2)	11 (52.4)
Education; *n* (%)		
Compulsory or middle school	3 (14.3)	6 (10.7)
Vocational training	12 (57.1)	33 (58.9)
High school	4 (19.0)	9 (16.1)
University	2 (9.5)	8 (14.3)
Days between Forms 1 and 2; *M* (*SD*)	n/a	60.95 (13.84)

Note. AWLT = Auditory Wordlist Learning Test; *M* = mean; *SD* = standard deviation.
[a]Including the 21 subjects who completed Forms 1 and 2.

Table 4. Reliability of the AWLT's Form 1.

	Reliability		
Measure	Odd–even split-half[a]	Internal consistency[b]	Parallel forms[c]
Learning phase			
Trial 1	0.49*	0.33	0.45*
Trial 2	0.62***	0.52	0.77***
Trial 3	0.68***	0.60	0.69**
Trial 4	0.73***	0.58	0.79***
Learning sum	0.95***	0.91	0.81***
Short delayed free recall	0.72***	0.66	0.82***
Long delayed free recall	0.78***	0.74	0.85***
Recognition accuracy	n/a	n/a	0.62**

Note. AWLT = Auditory Wordlist Learning Test.
[a]Pearson's *r* with Spearman–Brown correction for altered test length. [b]Cronbach's α. [c]Spearman's ρ.
*p < .05. **p < .01. ***p < .001.

Test Scores Within and Between Parallel Forms. Figure 2 displays the recall rates of the AWLT's first form at Trials 1 and 4 of the learning phase, as well as short and long delayed free recall. The curves showed the expected pattern. Recall rates increased from Trial 1 to Trial 4 and were lower in the long compared with the short delayed free recall. In Trials 1 and 4, the typical U shaped association between serial position and recall probability was apparent. For the delayed recalls, the recency effect was diminished and recall rates decreased with increasing serial position.

In general, mean scores of the AWLT in both forms showed the expected pattern (Figure 3). The mean number of recalled words increased from Trial 1 to Trial 4 of the learning phase and decreased at short and long delayed free recalled. Recognition accuracy was very high, as would be expected in a cognitively healthy sample.

Mean scores on the measures did not differ between the two forms, as could be expected from inspecting Figure 3. The degrees of freedom for *time* were corrected with the Greenhouse–Geisser formula to adjust for the violation of sphericity in the RM-ANOVA model. Mean scores aggregated across *form* increased from Trial 1 (*M* = 6.31, *SD* = 1.31) over Trials 2 (*M* = 8.55, *SD* = 1.73) and 3 (*M* = 9.21, *SD* = 1.80) to Trial 4 (*M* = 9.69, *SD* = 1.57) and decreased

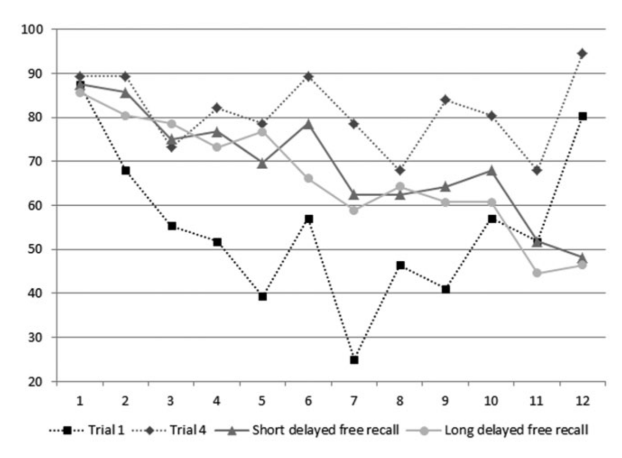


Figure 2. Recall rates for the 12 words of Form 1 at Trials 1 and 4 of the learning phase, as well as short and long delayed free recall.

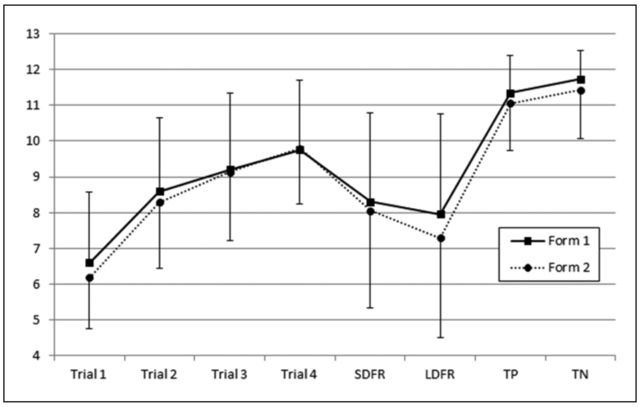


Figure 3. Mean scores and standard deviations of Forms 1 and 2 at Trials 1 to 4 of the learning phase, short delayed free recall (SDFR), long delayed free recall (LDFR), as well as true positives (TP) and true negatives (TN) at the recognition task.

at short (M = 7.91, SD = 2.55) and long delayed recall (M = 7.41, SD = 2.74), while the number of true positives was high (M = 11.02, SD = 1.18), F(2.27, 45.43) = 42.37, p < .001, partial η^2 = 0.68). There were no differences in scores between the two *forms* aggregated across *time*, F(1,20) = 0.17, p = .682, η2 = 0.01, and no differences between the forms over *time*, F(3.81, 76.29) = 1.18, p = .327, η2 = 0.06.

Linguistic Recall Effects in the AWLT. Spearman's ρ indicated no statistically significant correlations of the words' frequency and neighborhood size with their rates of immediate and delayed recall as well their recognition rates of the AWLT's first form. Frequency was not associated with recall rates at learning Trial 1 (ρ = 0.21, p = .333), learning Trial 4 (ρ = 0.28, p = .178), short delayed recall (ρ = 0.32, p = .123), long delayed recall (ρ = 0.01, p = .961), or recognition rate (ρ = 0.31, p = .136). Similarly, neighborhood size was not related to recall rates at learning Trial 1 (ρ = 0.01, p = .972), learning Trial 4 (ρ = −0.01, p = .981), short delayed recall (ρ = 0.11, p = .616), long delayed recall (ρ = 0.11, p = .594), or recognition rate (ρ = 0.05, p = .836). The results did not differ when considering both forms together or separately.

The following paragraphs report the results of the repeated measures analyses, which were performed with the test data of the AWLT's first form. The repeated measures refer to the recall scores at Trials 1 and 4 of the learning phase as well short and long delayed recall. That is, measures within one testing session, not between pilot Studies 1 and 2. As the parallel forms were found to be linguistically similar, linguistic recall effects were only examined in the first form. The degrees of freedom for *time* were corrected with the Greenhouse–Geisser formula due to the violation of sphericity for all RM-ANOVAs. Aggregated across all levels of *length*, recall rates increased from Trial 1 (M = 53.18, SD = 18.25) to Trial 4 (M = 80.75, SD = 17.70) and decreased again at short delayed (M = 70.14, SD = 22.29) and long delayed free recall (M = 67.86, SD = 23.63), F(2.27, 124.55) = 54.02, p < .001, partial η^2 = 0.50. *Length* had no influence on recall aggregated across *time,* F(1, 55) < .01, p = .984, partial η^2 < 0.01. Recall rates of short and long words differed over *time,* F(2.36, 129.93) = 4.00, p = .015, partial η^2 = 0.07.

Aggregated across all levels of *frequency*, recall rates increased from Trial 1 (M = 55.06, SD = 16.34) to Trial 4 (M = 81.25, SD = 16.31) and decreased again at short delayed (M = 69.20, SD = 20.65) and long delayed free recall (M = 66.37, SD = 2 3.30), F(2.16, 118.90) = 57.16, p < .001, partial η^2 = 0.51. Aggregated across *time*, high-frequency words (M = 70.46, SD = 18.14) were better recalled than low-frequency words (M = 65.48, SD = 19.89), F(1,55) = 4.84, p = .032, partial η^2 = 0.08). Recall rates differed between high- and low-frequency words over *time,* F(2.65, 145.64) = 4.02, p = .009, partial η^2 = 0.07.

Aggregated across all levels of *neighborhood*, recall rates increased from Trial 1 (M = 55.06, SD = 16.34) to Trial 4 (M = 81.25, SD = 16.31) and decreased again at short delayed (M = 69.20, SD = 20.65) and long delayed free recall (M = 66.37, SD = 23.30), F(2.16, 118.90) = 57.16, p < .001, partial η^2 = 0.51. Aggregated across time, *neighborhood* had no influence on recall rates, F(1, 55) = 0.11, p= .737, partial η^2 < 0.01. Recall rates differed between words with small and large neighborhood sizes over *time,* F(3, 165) = 5.45, p = .001, partial η^2 = 0.09.

The results suggest the expected difference in recall rates between the individual measures that demonstrate the AWLT's ability to measure learning and retention. The significant interaction effects indicate linguistic recall effects, which seemed to vary between the AWLT's measures. *Table 5* displays the decomposed interaction effects of linguistic factors with the time of recall in the AWLT. After adjusting for multiple comparisons with the Benjamini–Hochberg (1995) procedure only the recall advantage for high-frequency of low-frequency words at Trial 4 remained statistically significant with a Cohen's d of 0.48. A similar advantage for high-frequency words was found at Trial 1 (Cohen's d = 0.48; however, the association was not statistically significant.

Discussion

The AWLT showed good to very good internal consistency and reliability, especially for core variables like the learning sum, as well as short and long delayed recall. Test scores behaved as expected in cognitively healthy persons with an increase in the number of recalled words during the learning phase and a decrease at short and long delayed free recall, as well as good recognition performance. Despite the efforts to reduce linguistic recall effects by choosing words that

Table 5. Linguistic Recall Effects in the AWLT. Decomposition of the Significant Interaction Effects of Length, Frequency, And Neighborhood Size With The time of Recall in the AWLT.

	% Recalled				
Linguistic recall effects	*M* (*SD*)	Difference, *M* (*SD*)	Difference, 95% CI	*T*[a]	Cohen's *d*
Length					
Trial 1					
Short length	49.40 (29.81)	−7.54 (33.58)	[−16.53, 1.45]	−1.68	0.30
Long length	56.94 (18.48)				
Trial 4					
Short length	79.76 (23.51)	−1.98 (19.66)	[−7.25, 3.28]	−0.76	0.10
Long length	81.75 (16.33)				
SDFR					
Short length	72.02 (29.66)	3.77 (25.45)	[−3.05, 10.59]	1.11	0.14
Long length	68.25 (20.91)				
LDFR					
Short length	70.83 (29.17)	5.95 (26.38)	[−1.11, 13.02]	1.69	0.22
Long length	65.88 (24.78)				
Frequency					
Trial 1					
Low frequency	49.70 (21.43)	−10.71 (31.21)	[−19.07, −2.36]	−2.57	0.48
High frequency	60.42 (23.69)				
Trial 4					
Low frequency	76.49 (21.27)	−9.52 (22.89)	[−15.65, −3.39]	−3.11*	0.48
High frequency	86.01 (18.47)				
SDFR					
Low frequency	69.05 (25.71)	−0.30 (24.51)	[−6.86, 6.27]	−0.09	0.01
High frequency	69.35 (22.20)				
LDFR					
Low frequency	66.67 (25.62)	0.60 (23.13)	[−5.60, 6.79]	0.19	0.03
High frequency	66.07 (26.39)				
Neighborhood size					
Trial 1					
Small neighborhood	59.52 (23.11)	8.93 (32.72)	[0.17, 17.69]	2.04	0.20
Large neighborhood	50.59 (23.13)				
Trial 4					
Small neighborhood	83.93 (17.40)	5.36 (22.04)	[−0.55, 11.26]	1.82	0.27
Large neighborhood	78.57 (21.72)				
SDFR					
Small neighborhood	66.67 (23.99)	−5.06 (28.23)	[−12.62, 2.50]	−1.34	0.20
Large neighborhood	71.73 (26.00)				
LDFR					
Small neighborhood	63.40 (26.86)	−5.95 (26.10)	[−12.94, 1.04]	−1.71	0.23
Large neighborhood	69.35 (26.55)				

Note. SDFR = short delayed free recall; LDFR = long delayed free recall; AWLT = Auditory Wordlist Learning Test; *M* = mean; *SD* = standard deviation.
[a]Degrees of freedom = 55.
*Statistically significant according to the Benjamini–Hochberg procedure.

were linguistically as similar as possible, an advantage for words with high frequency was observed. Given that this effect has already been observed in the German CVLT (Hessler et al., 2016), which has similarly little variability with regard to word frequency, it can be concluded that the word frequency effect is prominent in wordlist learning tests and might not be fully preventable.

General Discussion

The AWLT is a new, reliable test of verbal learning, short- as well as long-term retention, and recognition. Evidence from psycholinguistic studies was used to increase the control over item and test difficulty, and the similarity between parallel forms as well as between learning and distraction lists. Linguistic recall effects that were found in the German CVLT (Hessler et al., 2016) could be reduced but not eliminated and an advantage for high frequency was still present in the AWLT.

With the AWLT, we aimed to balance a sufficient amount of diagnostic information with efficiency for the clinician and acceptability on behalf of the patients. The AWLT will be part of the neuropsychological test battery Cognitive Functions Dementia (CFD) for the early detection of dementia that is run on a tablet PC with connected external loudspeakers. To increase the standardization of word presentation, the AWLT's words were previously recorded in a studio and can be played to the test taker in the learning phase and the recognition task. In addition, it will be possible to record the answers in order to cross-check and, if necessary correct, the results after test completion. The present study employed pilot data from a paper-and-pencil version. In future studies, the AWLT's reliability, validity, and susceptibility to linguistic recall effects need to be investigated in its final tablet-based version.

Currently, norming and validation of the test battery are in progress so that norms for the AWLT will be available in 2017. As part of these efforts, established measures of verbal learning like the CVLT and the CERAD-WL are administered to both cognitively healthy persons and persons with mild cognitive impairment and dementia, allowing for the calculation of the AWLT's concurrent and construct validity, as well as other psychometric qualities based on data from a large sample. For now, the learning slope and recall rates found in the pilot study might serve as preliminary indicators of the AWLT's validity in assessing verbal memory.

The linguistic approach to word selection likely also benefits the development of new verbal memory tests in other languages. Effects like the preference for high-frequency words appear to occur at basic physiological levels of language processing (Diana & Reder, 2006; Inhoff & Rayner, 1986) and the theoretical background of the AWLT's construction is based on international studies. Employing similar linguistic construction principles, Italian and English versions of the AWLT are currently in development.

Linguistic Control of Test and Item Difficulty

The AWLT lies between CERAD-WL and CVLT not only with regard to its structure but also with regard to its linguistic profile. Experimental studies suggest that words with high frequency (MacLeod & Kampe, 1996) and large neighborhood sizes (Jalbert et al., 2011; Roodenrys et al., 2002) have higher recall rates. Given that the AWLT's words are more frequent and have larger neighborhood sizes than the CVLT's, the AWLT is likely easier than the CVLT not only due to the structure but also on the item level. While the average neighborhood size was similar, the AWLT's words were less frequent than the CERAD-WL's, suggesting that the AWLT's items are more difficult than the CERAD-WL's. Importantly, these theoretical considerations need to be confirmed by empirical evidence.

The AWLT is less extensive than the CVLT, as it has fewer words, a shorter learning phase, does not include a distraction learning list ("List B" in the CVLT), and does not include cued recall according to semantic categories. As a consequence, the AWLT produces fewer diagnostic variables (e.g., no score for semantic clustering) but likely is more tolerable to patients and might have higher completion rates in persons with cognitive impairment.

Parallel Forms and Reliability

Controlling the linguistic properties of the AWLT's words also ensures high parallelization between the two test forms and between learning and distraction lists. All four lists in the two forms are similar with regard to their words' length, frequency, neighborhood size, and semantic group membership. This high level of similarity up to the individual items is unique in the AWLT and increases the test's diagnostic accuracy in repeated testing and the recognition trial. In addition, mean scores did not differ between the two parallel forms.

Linguistic Recall Effects

The AWLT's variability within the linguistic variables was smaller compared with CERAD-WL and CVLT. Linguistic homogeneity reduces the likelihood of including words that have high linguistic salience, for example, due to a very high frequency compared with the rest of the list. By reducing the variability, the words are assumed to have similar linguistic recall properties. However, as in the CVLT (Hessler et al., 2016), high-frequency words had a small advantage during the learning phase of the AWLT. The effect was markedly smaller in the AWLT than in the CVLT, possibly due to the higher mean frequency of the words compared with the CVLT. Words with small neighborhoods had a very small advantage at Trial 1 of the AWLT. This effect was reversed in the CVLT, where cognitively healthy subjects better recalled words with large neighborhood sizes, as was also suggested by experimental studies (e.g., Jalbert et al., 2011; Roodenrys et al., 2002). Length had no influence on learning and recall in both tests. Even though small effects, especially pertaining to frequency, were present, the AWLT seemed to be fairly robust against linguistic recall effects. Yet, it remains to be investigated, how linguistic recall effects present in the AWLT in persons with cognitive impairment and whether differences between diagnostic groups exist.

Clinical Implications

The AWLT was primarily developed for the diagnostic use in late-life cognitive decline, but is also suited for assessing verbal memory in other contexts. Its words were chosen with the aim of increasing diagnostic fairness through reducing linguistic variability, as it has been proposed that linguistic memory effects may present differently depending on the test takers' cognitive status (Balota et al., 2002; Hessler et al., 2016, Soo-ampon et al., 2004; Wilson et al., 1983). The results from the present study suggest that in cognitively healthy subjects, effects like the advantage for high-frequency words, which was prominent in the German CVLT (Hessler et al., 2016), could be reduced but not completely eliminated in the AWLT. This comes as no surprise, as the AWLT showed similarly little variability in frequency between the words as the AWLT. Since there is no clinical data available yet, it remains to be investigated whether the AWLT is actually fairer than other tests of verbal memory. Ideally, these studies would examine the AWLT's feasibility, validity, and linguistic memory effects in a variability of patient groups, including, for example, aphasia, schizophrenia, and depression. Hypothetically, the AWLT might be better suited for patients that have retrieval difficulties than the CVLT, as the former's linguistic properties promote encoding and retrieval more than the latter's.

Conclusions

The AWLT is a reliable test for the assessment of learning and verbal memory. The application of psycholinguistic evidence in its construction allowed for higher control of item difficulty and better parallelization between forms as well as between learning and distraction lists. Even though word frequency affected learning performance, the AWLT seems to be less linguistically loaded.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: Evidence for two processes in episodic recognition performance. *Journal of Memory and Language*, *46*, 199–226.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289-300.

Butters, N. (1985). Alcoholic Korsakoff's syndrome: some unresolved issues concerning etiology, neuropathology, and cognitive deficits. *Journal of Clinical and Experimental Neuropsychology*, *7*, 181-210.

Criss, A. H., Aue, W. R., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language*, *64*, 119-132.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *Manual for the California Verbal Learning Test—Second Edition (CVLT-II)*. San Antonio, TX: Psychological Corporation.

Delis, D. C., Massman, P. J., Butters, N., Salmon, D. P., Cermak, L. S., & Kramer, J. H. (1991). Profiles of demented and amnesic patients on the California Verbal Learning Test: Implications for the assessment of memory disorders. *Psychological Assessment*, *3*, 19-26.

Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 805-815.

Geyken, A. (2007). The DWDS corpus: a reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Collocations and idioms: Linguistic, lexicographic, and computational aspects* (pp. 23-40). London, England: Continuum Press.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB—A lexical database for the psychological and linguistic research (article in German). *Psychologische Rundschau*, *62*, 10-20.

Hessler, J. B., Fischer, A. M., & Jahn, T. (2016). Differential linguistic recall effects in the German California Verbal Learning Test in healthy aging and Alzheimer's dementia: A retrospective analysis of routine diagnostic data. *Archives of Clinical Neuropsychology*, *31*, 1-11.

Hulme, C., Surprenant, A. M., Bireta, T. J., Stuart, G., & Neath, I. (2004). Abolishing the word-length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 98-106. doi:10.1037/0278-7393.30.1.98

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: effects of word frequency. *Attention, Perception, & Psychophysics*, *40*, 431-439.

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 338-353.

Lee, J. H., Lee, K. U., Lee, D. Y., Kim, K. W., Jhoo, J. H., Kim, J. H., & …Woo, J. I. (2002). Development of the Korean version of the Consortium to Establish a Registry for Alzheimer's Disease Assessment Packet (CERAD-K): clinical and neuropsychological assessment batteries. *Journal of Gerontology: Psychological Sciences, 57B*, 47–53.

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 132-142.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Morris, J. C., Heyman, A., Mohs, R. C., & Hughes, J. P. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, *39*, 1159-1165.

Niemann, H., Sturm, W., Thöne-Otto, A., & Willmes, K. (2008). *California Verbal Learning Test. Deutsche Adaptation*. Frankfurt am Main, Germany: Pearson Assessment & Information GmbH.

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1019-1034.

Soo-ampon, S., Wongwitdecha, N., Plasen, S., Hindmarch, I., & Boyle, J. (2004). Effects of word frequency on recall memory following lorazepam, alcohol, and lorazepam alcohol interaction in healthy volunteers. *Psychopharmacology*, *176*, 420-425.

Thalmann, B., & Monsch, A. U. (1997). *CERAD. The Consortium to Establish a Registry for Alzheimer's Disease. Neuropsychologische Testbatterie*. Basel, Switzerland: Memory Clinic Basel.

Wilson, R. S., Bacon, L. D., Kramer, R. L., Fox, J. H., & Kaszniak, A. W. (1983). Word frequency effect and recognition memory in dementia of the Alzheimer type. *Journal of Clinical Neuropsychology*, *5*, 97–104.

RE: Permission to include paper published in "Assessment" in dissertation

Michelle Binur <Michelle.Binur@sagepub.com> im Auftrag von permissions (US) <permissions@sagepub.com>

Mi 24.05.2017 19:53

Inbox

An: Heßler, Johannes <johannes.hessler@tum.de>;

Dear Johannes,

I work in the permissions department of SAGE Publishing. Thank you for your request.

You may use your article in your dissertation. Please note that this permission does not cover any 3rd party material that may be found within the work. You must properly credit the original source, *Assessment*. Please let us know if you have further questions.

Best regards,
Michelle Binur

Contract Administrator
SAGE Publishing
2455 Teller Road
Thousand Oaks, CA 91320
USA

www.sagepublishing.com
Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

From: Sabrina Buie
Sent: Wednesday, May 24, 2017 9:22 AM
To: permissions (US) <permissions@sagepub.com>
Subject: FW: Permission to include paper published in "Assessment" in dissertation

From: Authorqueries
Sent: Wednesday, May 24, 2017 9:20 AM
To: Heßler, Johannes <johannes.hessler@tum.de>
Cc: Sabrina Buie <Sabrina.Buie@sagepub.com>
Subject: RE: Permission to include paper published in "Assessment" in dissertation

Dear Johannes,

Thank you for your email to SAGE Publishing. I've copied-in the Senior Editorial Assistant, Sabrina Buie, who works on Assessment and will be able to assist you with your query.

Should you require my assistance in the future please don't hesitate to get in contact.

Best wishes,

Hannah

From: Heßler, Johannes [mailto:johannes.hessler@tum.de]
Sent: 22 May 2017 16:54
To: Authorqueries <Authorqueries@sagepub.com>
Subject: Permission to include paper published in "Assessment" in dissertation

Dear Madam or Sir,

my paper "Applying Psycholinguistic Evidence to the Construction of a New Test of Verbal Memory in Late-Life Cognitive Decline" was published in Assessment (online first). As the study described in the paper is part of my disertation project, I would like to include the full text in its published form in my dissertation. My university requires that every dissertation is uploaded to the library Website after graduation. Therefore, I wanted to request the permission of SAGE to include the paper in my doctoral thesis. Of course, I will indicate that the Copyright lies at SAGE and include the reference to the journal.

Thank you for your efforts an Kind regards
Johanens Hessler

Klinik und Poliklinik für Psychiatrie und Psychotherapie
Klinikum rechts der Isar der TU München
Ismaninger Straße 22, 81675 München

Fon: +49 89 4140 4209
Mail: johannes.hessler@tum.de

Appendix D

Linguistic fairness and differential validity of the Auditory Wordlist Learning Test (AWLT) in dementia of the Alzheimer’s type

Linguistische Fairness und differentielle Validität des Auditiven Wortlisten Lerntests (AWLT) bei Demenz vom Alzheimer-Typ

Linguistic fairness and differential validity of the Auditory Wordlist Learning Test (AWLT) in dementia of the Alzheimer's type

Johannes Baltasar Hessler,[1] David Brieber,[2] Johanna Egle,[2] Georg Mandler,[2] Thomas Jahn[1]

[1] Klinik und Poliklinik für Psychiatrie und Psychotherapie, Klinikum rechts der Isar der Technischen Universität München, Deutschland

[2] Schuhfried GmbH, Mödling, Österreich

Korrespondenz:

Johannes Baltasar Hessler
Klinik und Poliklinik für Psychiatrie und Psychotherapie
Klinikum rechts der Isar der Technischen Universität München
Ismaninger Straße 22
81675 München
johannes.hessler@tum.de

Kurztitel: Auditiver Wortlisten Lerntest (AWLT)

Zusammenfassung

Der Auditive Wortlisten Lerntest (AWLT) ist Teil des Test-Sets Kognitive Funktionen Demenz (CFD; Cognitive Functions Dementia) im Rahmen des Wiener Testsystems (WTS). Der AWLT wurde entlang neurolinguistischer Kriterien entwickelt, um Interaktionen zwischen dem kognitiven Status der Testpersonen und den linguistischen Eigenschaften der Lernliste zu reduzieren. Anhand einer nach Alter, Bildung und Geschlecht parallelisierten Stichprobe von gesunden Probanden (N=25) und Patienten mit Alzheimer Demenz (N=25) wurde überprüft, inwieweit dieses Konstruktionsziel erreicht wurde. Weiter wurde die Fähigkeit der Hauptvariablen des AWLT untersucht, zwischen diesen Gruppen zu unterscheiden. Es traten linguistische Gedächtniseffekte auf, jedoch keine Interaktionen mit den diagnostischen Gruppen. Die Hauptvariablen trennten mit großen Effektstärken Patienten von Gesunden. Der AWLT scheint bei vergleichbarer differenzieller Validität linguistisch fairer als vergleichbare Instrumente zu sein.

Schlüsselwörter: Verbalgedächtnis, Testentwicklung, Alzheimer Demenz, Validität

Abstract

The Auditory Wordlist Learning Test is part of the test-set Cognitive Functions Dementia (CFD) for the Vienna Test System (VTS). The AWLT was developed along neurolinguistic criteria to prevent interactions between the cognitive status of the test-persons and the linguistic characteristics of the learning list. With data from a sample of healthy persons (N=25) and persons with Alzheimer's dementia (N=25) who were parallelized according to age, education, and sex, we investigated whether this aim of test construction was met. We found linguistic recall effects, however, no interactions with the diagnostic group. The AWLT's main variables differentiated with large effect sizes between the groups. The AWLT seems to be linguistically fairer and equally valid compared to similar instruments.

Key words: verbal memory, test development, Alzheimer's dementia, validity

Einleitung

Störungen des Verbalgedächtnisses sind das Kardinalsymptom der häufigen Demenz vom Alzheimer-Typ (DAT) (Bondi et al., 2008). Sie treten meist schon früh im Krankheitsverlauf auf (Bäckman, Jones, Berger, Laukka, & Small, 2005) und besitzen eine hohe prognostische Validität für die Progression von der Leichten Kognitiven Beeinträchtigung (mild cognitive impairment; MCI) zur Demenz (Gainotti, Quaranta, Vita, & Marra, 2014). Ein genaue Differenzierung, welche Aspekte des Verbalgedächtnisses beeinträchtigt sind, kann die Abgrenzung von DAT zu Depression (Jahn et al., 2004) und Frontotemporaler Demenz (Diehl & Kurz, 2002) unterstützen.

Als Teil der Touchscreen-basierten neuropsychologischen Testbatterie zur Demenzdiagnostik, dem Test-Set „Kognitive Funktionen Demenz“ (Jahn & Hessler, 2017) im Rahmen des Wiener Testsystems (WTS), wurde der Auditive Wortlisten Lerntest (Hessler & Jahn, 2017) zur Erfassung verschiedener Aspekte des episodischen Verbalgedächtnisses konzipiert. In der Normstichprobe des CFD fanden sich für sämtliche Kennwerte des AWLT ausreichend große Reliabilitäten (Jahn & Hessler, 2017). Der AWLT besteht aus vier Testphasen (Lernen, kurz verzögerter freier Abruf, lang verzögerter freier Abruf und Wiedererkennen), die zusammen mit den vorgeschriebenen Pausen etwa 40 Minuten dauern. *Tabelle I* im elektronischen Supplement zeigt den Ablauf des AWLT.

Der AWLT basiert auf dem bekannten Wortlistenlernparadigma, bei dem den Probandinnen und Probanden wiederholt eine Reihe von Wörtern präsentiert wird, die unmittelbar und nach Verzögerung frei erinnert sowie unter ähnlichen Distraktoren wiedererkannt werden sollen. Neu ist beim AWLT, dass er als Teil des Test-Sets CFD vollständig am Touchscreen durchgeführt werden kann. Dies ermöglicht, die Wörter standardisiert durch den Computer vorlesen zu lassen, die Antworten der Probandinnen und Probanden aufzunehmen und später mit den während der Testdurchführung kodierten Antworten (richtiges Wort, Intrusion, Wortwiederholung) abzugleichen. Außerdem sind auch kompliziertere Kennwerte sofort verfügbar, beispielsweise der Serielle Cluster Index (Stricker, Brown, Wixted, Baldo, & Delis, 2002), der auf der Reihenfolge basiert, in der die

Wörter erinnert werden. Zudem beruhen die Konstruktionsprinzipien des AWLT auf psycholinguistischer Evidenz aus klinischen und experimentellen Studien, die in der Entwicklung anderer Wortlisten Lerntests zwar ebenfalls eine Rolle spielten, beispielsweise bei der Weiterentwicklung des originalen California Verbal Learning Test (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000), jedoch nicht im gleichen Umfang wie beim AWLT (Hessler, Brieber, Egle, Mandler, & Jahn, 2017).

Verschiedene Untersuchungen deuten darauf hin, dass sich die linguistischen Eigenschaften des Lernmaterials eines Wortlisten-Lerntests auf die Testleistung der Probanden auswirken können. Beispielsweise werden Wörter, die in einer Sprache häufiger vorkommen, im Vergleich zu seltenen Wörtern besser gelernt und erinnert (MacLeod & Kampe, 1996). Manche linguistischen Worteigenschaften scheinen zudem mit dem kognitiven Status der Testpersonen zu interagieren, zum Beispiel bei der orthografischen Nachbarschaftsgröße. Ein orthografischer Nachbar ist definiert als jedes sinnvolle Wort, das sich durch Addition, Subtraktion oder Substitution eines Buchstabens eines bestimmten Wortes ergibt, wobei pro Nachbar nur eine Operation vorgenommen werden kann (Jalbert, Neath, & Surprenant, 2011a). Während in der deutschen Adaptation des CVLT (Niemann, Sturm, Thöne-Otto, & Willmes, 2008) ältere Personen ohne kognitive Beeinträchtigung Wörter mit vielen orthografischen Nachbarn besser erinnerten, zeigten Personen mit DAT eine Präferenz für Wörter mit wenigen Nachbarn (Hessler, Fischer, & Jahn, 2016). Ähnliche Effekte werden auch für die Worthäufigkeit berichtet (Hessler et al., 2016; Soo-ampon, Wongwitdecha, Plasen, Hindmarch, & Boyle, 2004; Wilson, Bacon, Fox, Kramer, & Kaszniak, 2008). Diese Effekte könnten die Fairness und damit die diagnostische Genauigkeit eines Tests einschränken, in dem die linguistischen Eigenschaften der zu lernenden Wortliste für bestimmte Personen einen Lern- und Abrufvorteil ergeben, für andere aber nicht (Hessler et al., 2016; 2017).

Ausgehend von der Annahme, dass die linguistischen Eigenschaften einer Wortlernliste für den Lernerfolg und die Behaltensleistung relevant sind, wurden die Wörter des AWLT hinsichtlich ihrer orthografischen Länge (Silbenzahl), ihrer Häufigkeit in der deutschen Sprache und der orthografischen Nachbarschaftsgröße ausgewählt. Um die Ähnlichkeit der Parallelformen zu erhöhen, wurden die Wörter so ausgewählt, dass sich die jeweiligen Mittelwerte der linguistischen Variablen nicht unterscheiden. Anhand der Anzahl der zu lernenden Worte und der Anzahl der Lerndurchgänge wurde zudem dafür gesorgt, dass sich der AWLT hinsichtlich seiner Schwierigkeit zwischen der Wortliste der deutschsprachigen Adaptation der neuropsychologischen Testbatterie des Consortium to Establish a Registry for Alzheimer's Disease (CERAD-NTB; Thalmann & Monsch, 1997) und dem deutschen CVLT positioniert. Ein drittes Konstruktionsprinzip war, die Variabilität von Wortlänge, Worthäufigkeit und Nachbarschaftsgröße so klein wie möglich zu halten, um Interaktionen mit dem kognitiven Status der Patienten vorzubeugen. So sollten Effekte wie der Vorteil für Wörter mit relativ vielen Nachbarn für kognitiv Gesunde und der Vorteil für Wörter mit relativ wenig Nachbarn für Patienten mit DAT verhindert werden. Bisher wurde jedoch nicht untersucht, ob dieses dritte Konstruktionsziel der linguistischen Fairness mit dem AWLT erreicht wurde. In einer Pilotstudie mit kognitiv Gesunden wurde ein leichter Vorteil für häufige Wörter in der Lernphase gefunden (Hessler et al., 2017).

Der AWLT gibt nach vollständiger Durchführung 33 Kennwerte aus, darunter vier Haupt-, 13 Neben- und 16 Zusatzvariablen. Zusätzlich wird im CFD ein Gesamtindex für den AWLT berechnet, in den die vier Hauptvariablen gewichtet nach ihrer Ladung in einem Strukturgleichungsmodell der CFD-Kennwerte eingehen (Jahn & Hessler, 2017). Da sich der AWLT hinsichtlich des ihm zugrunde liegenden Untersuchungsparadigmas nicht von anderen Wortlisten-Lerntests unterscheidet, ist davon auszugehen, dass seine Kennwerte mindestens ebenso valide zwischen Personen mit und ohne DAT unterscheiden. Aus diesem Grund lag das Hauptaugenmerk der hier berichteten Arbeit darauf, wie sich die linguistischen Merkmale des AWLT in einer großen Stichprobe aus kognitiv Gesunden sowie kognitiv Beeinträchtigten darstellen.

Basierend auf der Normierungsstichprobe des CFD und einer ersten klinischen Validierungsstichprobe, die Personen mit DAT einschließt, untersucht die vorliegende Studie zwei Fragen: (1) Treten im AWLT Interaktionen zwischen den linguistischen Eigenschaften seines Lernmaterials und dem kognitiven Status der Testpersonen auf? (2) Wie gut unterscheiden die Hauptvariablen des AWLT sowie der Index Verbales Langzeitgedächtnis zwischen gesunden Probanden und Personen mit DAT?

Methode

Stichprobe

Die Daten für die vorliegende Arbeit stammen aus der deutsch-österreichischen Normierungsstudie des CFD sowie aus einer multizentrischen klinischen Validierungsstudie. Das Studienprotokoll für die klinische Validierung des CFD wurde durch die Ethikkommission des Klinikums rechts der Isar der Technischen Universität München ohne Einschränkungen genehmigt (Az. 353/16 S, 08.08.2016). Für die vorliegende Veröffentlichung wurden aus dem Datensatz der derzeit (Sommer 2017) noch laufenden klinischen Validierungsstudie alle Patientinnen und Patienten mit DAT ausgewählt (N = 25). Jedem dieser Patientinnen und Patienten wurde eine hinsichtlich Alter, Geschlecht und Bildungsgrad vergleichbare Person aus der Normierungsstichprobe zugeordnet, sodass die beiden Gruppen hinsichtlich dieser Merkmale exakt übereinstimmen.

Material und Procedere

Der AWLT wurde, wie drei andere Tests im CFD auch (VISCO, WIWO, WOBT), zwar speziell für dieses neue Test-Set entwickelt, kann im weiteren Rahmen des Wiener Testsystems aber auch als eigenständiger Test bei anderen Fragestellungen als der Demenzdiagnostik eingesetzt werden. Hierfür steht eine repräsentative Normstichprobe aus 168 (54 %) Frauen und 144 (46 %) Männern im Alter von 18 bis 94 Jahren (*M* = 49.04; *SD* = 18.64) zur Verfügung. Speziell der Demenzdiagnostik dient die „Normstichprobe 50+ CFD“ aus 163 (40 %) Männern und 244 (60 %) Frauen im Alter von 50 bis 94 Jahren (*M* = 67.80; *SD* = 9.95) (jeweils S1-Version), anhand der sämtliche Haupt- und Nebenvariablen des CFD konormiert wurden.

Der AWLT erfasst verschiedene Aspekte der verbalen Lern- und Merkfähigkeit und wurde im Rahmen der Validierung vollständig auf dem Touchscreen durchgeführt. In der Lernphase und der Wiedererkennung wurden die Wörter standardisiert durch den Computer vorgelesen. Vor dem ersten Lerndurchgang wurde ein Hinweis auf die nun folgenden Wörter abgespielt, um die Probanden an die Computerstimme zu gewöhnen und sicherzugehen, dass die Wörter laut genug abgespielt und verstanden werden. Die vier Hauptvariablen und der Gesamtindex des AWLT sind in *Tabelle 1* beschrieben.

TABELLE 1 UNGEFÄHR HIER

Statistische Analyse

Linguistische Fairness

Der Einfluss von Wortlänge, Worthäufigkeit und Nachbarschaftsgröße auf die Abrufraten in Lernen und verzögertem Abruf im AWLT wurden mittels Varianzanalysen mit Messwiederholungen (repeated measures analysis of variance; RM-ANOVA) untersucht (Hessler et al., 2016; 2017). Zu diesem Zweck wurde die Lernliste des AWLT für jedes linguistische Merkmal am Median dichotomisiert, sodass je zwei Gruppen entstehen (kurze versus lange Wörter; seltene versus häufige Wörter; Wörter mit wenigen versus vielen Nachbarn). Innerhalb jedes Merkmals können diese beiden Gruppen als experimentelles Treatment verstanden werden, dem die Probanden bei der Durchführung des AWLT ausgesetzt werden.

Für die Lerndurchgänge 1 und 4 sowie den lang verzögerten Abruf wurden anschließend die Abrufraten der Wörter in Abhängigkeit ihrer linguistischen Ausprägung als abhängige Variable berechnet. So wurde beispielsweise jeweils die Prozentzahl der bei Lerndurchgang 1 erinnerten kurzen und langen Wörter ermittelt. Für jedes linguistische Merkmal wurde ein 2 × 3 × 2 RM-ANOVA-Model mit dem Zwischensubjektfaktor *Diagnose* (Unbeeinträchtigt versus DAT) und den Innersubjektfaktoren *Subtest* (Lerndurchgang 1 versus Lerndurchgang 4 versus lang verzögerter Abruf) und *Linguistisches Merkmal* (niedrig versus hoch) mit den Abrufraten als abhängige Variablen erstellt.

Da der AWLT auf dem bewährten Wortlisten Lernparadigma basiert, war davon auszugehen, dass sich statistisch signifikante Haupteffekte für *Diagnose* und *Subtest* sowie eine Interaktion zwischen den beiden Faktoren finden werden. Aus diesem Grund wurden nur Haupt- und Interaktionseffekte, die eine linguistische Variable beinhalten mit post-hoc *t*-Tests weiter untersucht. Für post-hoc Tests wurde das Signifikanzniveau nach der Bonferroni-Methode adjustiert (α = 0.05/Anzahl Tests). Zudem wurden Effektstärken sowohl für die Effekte der RM-ANOVAs (partial η^2) als auch für die post-hoc *t*-Tests (Cohens *d*) berechnet.

Differentielle Validität

Zur Überprüfung der Validität des AWLT wurden die vier Hauptvariablen sowie der Gesamtindex auf ihre Fähigkeit untersucht, zwischen Unbeeinträchtigten und Personen mit DAT zu unterscheiden. Für jeden Kennwert wurde ein Receiver-Operator-Characteristics Analyse gerechnet, die die Area Under the Curve (AUC) als Hauptstatistik ausgibt, sowie Cohen’s *d* als Schätzer der Effektgröße der Mittelwertunterschiede. Alle statistischen Analysen wurden mit SPSS 24 für Microsoft Windows durchgeführt.

Ergebnisse

In der klinischen Validierungsstichprobe des CFD befanden sich zum Stichtag 25 Patient mit DAT, sodass die hier zugrunde gelegte Gesamtstichprobe aus 50 Personen bestand. In jeder Gruppe befanden sich 13 (52.0 %) Männer; das durchschnittliche Alter lag jeweils bei 72,44 (*SD* = 9.11) Jahren mit einer Streubreite von 54 bis 85. Pro Gruppe hatten 6 (24.0 %) Personen einen Haupt- oder Realschulabschluss ohne anschließende Berufsausbildung, 12 (48.0 %) eine abgeschlossene Berufsausbildung, 2 (8.0 %) eine höhere Schule mit Abitur oder Matura abgeschlossen und 5 (20.0 %) einen Universitäts- oder Hochschulabschluss. Für die Variable Wiedererkennen und den Index Verbales

Langzeitgedächtnis lagen bei zwei Patienten mit DAT keine Werte vor. Diese Patienten wurden jedoch nicht ausgeschlossen, um den Datensatz nicht weiter zu verkleinern. Diese fehlenden Werte wirken sich nur auf die Analysen der differentiellen Validität für die Variable Wiedererkennen und den Index Verbales Langezeitgedächtnis aus.

Linguistische Fairness

Aus Gründen der Übersicht werden im Folgenden nur statistisch signifikante Haupteffekte linguistischer Variablen detailliert beschrieben. Alle Ergebnisse der RM-ANOVAs sind in *Tabelle 2* dargestellt. Wortlänge, Worthäufigkeit und Nachbarschaftsgröße hatten einen Einfluss auf die Abrufraten über alle Subtests und Diagnosegruppen hinweg. Kurze Wörter hatten niedrigere Abrufraten (Gesamtmittelwert = 33.78, SD = 25.98) als lange Wörter (Gesamtmittelwert = 38.96, *SD* = 19.89), F ($df_{\text{Wortlänge}}$, df_{Fehler}) = 4.17 (1, 48), p = 0.048, partial η^2 = 0.08. Seltene Wörter hatten niedrigere Abrufraten (Gesamtmittelwert = 30.00, *SD* = 24.62) als häufige (Gesamtmittelwert = 40.22, *SD* = 20.40), F ($df_{\text{Worthäufigkeit}}$, df_{Fehler}) = 16.36 (1, 48), $p < 0.001$, partial η^2 = 0.25. Wörter mit wenigen Nachbarn hatten niedrigere Abrufraten (Gesamtmittelwert = 32.50, *SD* = 20.57) als Wörter mit vielen Nachbarn (Gesamtmittelwert = 40.25, *SD* = 21.82), F ($df_{\text{Wortlänge}}$, df_{Fehler}) = 13.06 (1, 48), p = 0.001, partial η^2 = 0.21.

TABELLE 2 UNGEFÄHR HIER

Die post-hoc *t*-Tests für die beiden signifikanten Interaktionseffekte sind in *Tabelle 3* beschrieben. In Lerndurchgang 1 bestand ein Vorteil von langen gegenüber kurzen Wörtern, während in Lerndurchgang 4 und dem lang verzögerten Abruf kein Unterschied gefunden wurde. Häufige Wörter wurden besser in den Lerndurchgängen 1 und 4 erinnert, während im lang verzögerten Abruf kein Unterschied bestand. Die *Abbildung 1* zeigt die Abrufraten der Wörter in Abhängigkeit der linguistischen Variablen, der Diagnose der Probanden und dem Subtest.

TABELLE 3 UNGEFÄHR HIER

ABBILDUNG 1 UNGEFÄHR HIER

Differentielle Validität

Die Hauptvariablen sowie der Indexwert zeigten eine hohe differentielle Validität für die Unterscheidung zwischen Unbeeinträchtigten und Patienten mit DAT. Die Effektstärken (Cohens *d*) für die Gruppenunterschiede waren größer als −1.50 und die AUCs lagen sämtlich über 0.85. Der kurz verzögerte Abruf unterschied am besten zwischen den Gruppen, gefolgt von dem Gesamtindex, dem lang verzögerten Gedächtnisabruf, dem Wiedererkennen und der Lernsumme. Die Ergebnisse sind in *Tabelle 4* dargestellt.

TABELLE 4 UNGEFÄHR HIER

Diskussion

Der AWLT ist ein Touchscreen-gestützter Test zur Erfassung der verbalen Lern- und Merkfähigkeit nach dem bewährten Wortlisten-Lernparadigma, der wie hier gezeigt mit großen Effektstärken zwischen älteren Personen mit und ohne DAT unterscheiden kann. Das Konstruktionsziel der linguistischen Fairness, also die Reduktion oder bestenfalls Verhinderung von Interaktionen zwischen dem kognitiven Status von Testpersonen und den Worteigenschaften der Lernliste, scheint zumindest für Personen mit DAT erreicht.

Die vorliegende Studie bestätigt die Ergebnisse aus der Pilotstudie zum AWLT (Hessler et al., 2017), die darauf hindeuten, dass linguistische Gedächtniseffekte in Wortlisten-Lerntests kaum vermieden werden können. Unabhängig von Diagnose und Subtest waren in der vorliegenden Studie höhere Silbenzahl, häufigeres Auftreten in der Sprache und größere Nachbarschaft mit erhöhten Abrufraten assoziiert. Diese Zusammenhänge treten im AWLT vor allem in den Lerndurchgängen auf. Während sich die Effekte von Häufigkeit und Nachbarschaftsgröße mit den Ergebnissen aus experimentelle

Studien decken (Jalbert et al., 2011a; MacLeod & Kampe, 1996), werden in gemischten Listen mit sowohl kurzen als auch langen Wörtern üblicherweise keine Effekte für die Wortlänge gefunden (Bireta, Neath, & Surprenant, 2006; Jalbert, Neath, Bireta, & Surprenant, 2011b).

Ein Vergleich mit Ergebnissen für den deutschen CVLT aus einer methodisch ähnlichen Studie, die im Vorfeld der Entwicklung des AWLT durchgeführt wurde (Hessler et al., 2016), ist in *Tabelle 6* abgebildet. In beiden Studien wurden kognitiv unbeeinträchtigte Personen mit Patienten mit DAT verglichen. Die linguistischen Analysen wurden nach der gleichen Methode mit RM-ANOVAs für den ersten und letzten Lerndurchgang sowie den lang verzögerten Abruf ausgeführt. Während im CVLT kein Einfluss der Wortlänge auf den Abruf gefunden wurde, zeigte der AWLT einen entsprechenden Haupteffekt und einen Interaktionseffekt mit den Subtests. CVLT und AWLT waren ähnlich darin, dass für beide Tests sowohl ein Haupteffekt der Worthäufigkeit als auch eine Interaktion mit den Subtests gefunden wurde. Bei der Nachbarschaftsgröße wies der CVLT eine Interaktion mit der Diagnose auf und der AWLT einen Haupteffekt. Im AWLT wurden somit fünf Effekte mit linguistischer Beteiligung gefunden, im CVLT vier. Weiter betreffen diese Effekte im AWLT alle untersuchten linguistischen Variablen, im CVLT nur Worthäufigkeit und Nachbarschaftsgröße. Betrachtet man die Variablen einzeln, scheint der AWLT jedoch nur bei der Wortlänge im Nachteil. Bei der Worthäufigkeit fällt der Haupteffekt im CVLT deutlich stärker aus als im AWLT, während die Interaktionseffekte klein (CVLT) bzw. mittelgroß (AWLT) sind. Der Interaktionseffekt im CVLT zwischen Nachbarschaftsgröße und Diagnose erreichte eine große Effektstärke und wiegt dazu schwerer als der Haupteffekt im AWLT (mittlere Effektstärke), da er potentiell die linguistische Fairness und Validität des Instruments beeinträchtigen kann. Wichtig ist hier zu bemerken, dass die Interaktion zwischen Diagnose und Nachbarschaftsgröße im AWLT zwar nur knapp die statistische Signifikanz verfehlte, jedoch eine kleine Effektstärke aufwies. Insgesamt deuten die Ergebnisse darauf hin, dass der AWLT somit linguistisch fairer zu sein als der deutsche CVLT.

Linguistische Interferenz ist kaum vollständig aus Wortlisten Lerntests zu eliminieren. Umso wichtiger ist es, sich schon bei der Testentwicklung mit diesem Aspekt zu beschäftigen und die Wörter dementsprechend auszuwählen. Bestehende Tests sollten hinsichtlich ihrer linguistischen Fairness untersucht werden, um etwaige Auswirkungen auf die Validität abschätzen und diskutieren zu können.

Die differentielle Validität des AWLT ist wenig überraschend, musste aber nichtsdestoweniger demonstriert werden. Der AWLT basiert auf dem klassischen Wortlisten-Lernparadigma, das zu den gebräuchlichsten und aussagekräftigsten Testverfahren in der neuropsychologischen Demenzdiagnostik gehört. Der AWLT zeigt ähnliche Effektstärken wie andere Tests der verbalen Lern- und Merkfähigkeit für die Unterscheidung zwischen älteren Personen ohne kognitive Beeinträchtigung und solchen mit MCI oder Demenz (Han, Nguyen, Stricker, & Nation, 2017; Jahn & Werheid, 2015; Zakzanis, Leach, & Kaplan, 1999). Während in der vorliegenden Untersuchung alle Kennwerte große Effektstärken aufwiesen, scheinen der kurz und lang verzögerte freie Gedächtnisabruf sowie der Index Verbales Langzeitgedächtnis besonders für die Unterscheidung Gesund versus DAT geeignet. Inwiefern die diagnostische Brauchbarkeit des AWLT beeinflusst ist durch den linguistischen Konstruktionsansatz und die standardisierte sowie digitalisierte Vorgabe und Durchführung, muss an dieser Stelle noch offen bleiben. In zukünftigen Studien sollte untersucht werden, ob sich die Eignung des AWLT zur Früherkennung und Differenzialdiagnose demenzieller Syndrome von derjenigen anderer Wortlisten Lerntests, beispielswiese des CVLT, in denselben Stichproben unterscheidet.

Neben dem CVLT bieten nur der AWLT und mit Einschränkung der verbale Lern- und Merkfähigskeitstest (VLMT; Helmstaedter, Lendt, & Lux, 2001) eine vergleichbare Vielzahl an Kennwerten zur Quantifizierung unterschiedlicher Teilaspekte verbaler Lern- und Gedächtnisprozesse, die für die klinische Differenzialdiagnostik mnestischer Störungen erwiesenermaßen nützlich sind oder sich noch als nützlich erweisen könnten (Niemann et al., 2008). Naheliegend ist auch der Vergleich des AWLT mit der Wortliste der CERAD-NTB, da beide Wortlisten Lerntests Teil einer Testbatterie zur neuropsychologischen

Demenzdiagnostik sind. Im Gegensatz zu den insgesamt 33 (bislang allerdings noch nicht vollzählig normierten) Ergebnisvariablen des AWLT schöpft die CERAD-Wortliste mit ihren nur acht Ergebnisvariablen ihr diagnostisches Potential nicht aus. Zum Beispiel werden beim Rekognitionsversuch (Wiedererkennen) die Hits als Anzahl der richtig-positiven Reaktionen nicht mit der Anzahl falsch-positiver Reaktionen kontrastiert. Viele Hits können folglich aus unterschiedlichem Antwortverhalten entstehen. Kognitiv Unbeeinträchtigte erkennen die gelernten Wörter in der Regel mit hoher Genauigkeit wieder, haben also viel Hits. Patienten mit Demenz dagegen entwickeln oft eine Tendenz zum „Ja-Sagen“ und erzielen dadurch viele Hits bei zugleich jedoch vielen falsch-positiven Reaktionen. Der AWLT gibt neben einem Diskriminabilitätsmaß der Signalentdeckungstheorie (Hauptvariable Wiedererkennen) die Hits, die Falsch-Positiven sowie die Ja-Sage-Tendenz aus. Einige Kennwerte des AWLT sind vom CVLT entlehnt, darunter der serielle Cluster-Index (Stricker et al., 2002) und die Abruf-Diskriminabilität (Delis et al., 2005). Einzigartig für den AWLT ist der aus dem Strukturgleichungsmodell für das Test-Set CFD stammende Indexwert, der die vier Hauptvariablen gewichtet kombiniert und somit auf einen Blick eine zusammenfassende und dennoch diagnostisch aussagekräftige Einschätzung des globalen Schweregrades mnestischer Defizite liefert. Andere Kennwerte, zum Beispiel zu Primacy- und Recency-Tendenzen beim Abruf, wurden neu entwickelt und müssen noch auf ihre klinische und diagnostische Relevanz überprüft werden.

Die vorliegende Studie ist limitiert durch die vorerst noch relativ geringe Zahl an Patienten mit DAT, die im weiteren Verlauf der klinischen Validierungsstudie erhöht werden wird. Angesichts der Ergebnisse dieser Zwischenauswertung ist jedoch davon auszugehen, dass Analysen von größeren Stichproben die linguistische Fairness und differenzielle Validität des AWLT weiter bestätigen werden.

Danksagung

Wir bedanken uns bei allen Testleitern in München und Wien für die Erhebung der Normstichprobe sowie bei unseren Kooperationspartnern (AMEOS Klinikum Bernburg, kbo-

Inn-Salzach Klinikum Wasserburg, Klinikum Altenburger Land, Klinikum Leer, Zentrum für Kognitive Störungen am Klinikum rechts der Isar der Technischen Universität München, Kolpinghaus Leopoldstadt, Praxisgemeinschaft Ambulante Neuropsychologische Psychiatrische Rehabilitation Aachen (ANPRA), Schön Klinik München Schwabing, Städtisches Klinikum Dessau) für die Erhebung der klinischen Daten.

Literatur

Bäckman, L., Jones, S., Berger, A.-K., Laukka, E. J., & Small, B. J. (2005). Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. *Neuropsychology*, *14*, 520–531.

Bireta, T. J., Neath, I., & Surprenant, A. M. (2006). The syllable-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review*, *13*, 434–438.

Bondi, M. W., Jak, A. J., Delano-Wood, L., Jacobson, M. W., Delis, D. C., & Salmon, D. P. (2008). Neuropsychological contributions to the early identification of Alzheimer's disease. *Neuropsychology Review*, *18*, 73–90.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *CVLT-II: California Verbal Learning Test: adult version*. San Antonia: Psychological Corporation.

Delis, D. C., Wetter, S. R., Jacobson, M. W., Peavy, G., Hamilton, J., Gongvatana, A., et al. (2005). Recall discriminability: utility of a new CVLT-II measure in the differential diagnosis of dementia. *Journal of the International Neuropsychological Society*, *11*, 708–715.

Diehl, J., & Kurz, A. (2002). Frontotemporal dementia: patient characteristics, cognition, and behaviour. *International Journal of Geriatric Psychiatry*, *17*, 914–918.

Gainotti, G., Quaranta, D., Vita, M. G., & Marra, C. (2014). Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease. *Journal of Alzheimer's Disease*, *38*, 481–495.

Han, S. D., Nguyen, C. P., Stricker, N. H., & Nation, D. A. (2017). Detectable neuropsychological differences in early preclinical Alzheimer's disease: a meta-analysis, 1–21, online first.

Helmstaedter, C., Lendt, M., & Lux, S. (2001). *VLMT - Verbaler Lern-und Merkfähigkeitstest*. Göttingen: Beltz Test GmbH.

Hessler, J. B., & Jahn, T. (2017). *Manual Auditiver Wortlisten Lerntest (AWLT)*. Mödling: Schuhfried GmbH.

Hessler, J. B., Brieber, D., Egle, J., Mandler, G., & Jahn, T. (2017). Applying Psycholinguistic

Evidence to the Construction of a New Test of Verbal Memory in Late-Life Cognitive Decline. *Assessment*, online first.

Hessler, J. B., Fischer, A. M., & Jahn, T. (2016). Differential Linguistic Recall Effects in the California Verbal Learning Test in Healthy Aging and Alzheimer's Dementia: Analysis of Routine Diagnostic Data. *Archives of Clinical Neuropsychology*, *31*(7), 689–699.

Jahn, T., & Hessler, J. B. (2017). *Handanweisung Kognitive Funktionen Demenz (CFD)*. Mödling: Schuhfried GmbH.

Jahn, T., & Werheid, K. (2015). Demenzen. Göttingen: Hogrefe.

Jahn, T., Theml, T., Diehl, J., Grimmer, T., Heldmann, B., Pohl, C., et al. (2004). CERAD-NP und Flexible Battery Approach in der neuropsychologischen Differenzialdiagnostik Demenz versus Depression. *Zeitschrift Für Gerontopsychologie & -Psychiatrie*, *17*, 77–95.

Jalbert, A., Neath, I., & Surprenant, A. M. (2011a). Does length or neighborhood size cause the word length effect? *Memory & Cognition*, *39*, 1198–1210.

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. E. M. (2011b). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 338–353.

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 132–142.

Niemann, H., Sturm, W., Thöne-Otto, A. I., & Willmes, K. (2008). *CVLT: California Verbal Learning Test: deutschsprachige Adaption: Manual*. Frankfurt am Main: Pearson.

Soo-ampon, S., Wongwitdecha, N., Plasen, S., Hindmarch, I., & Boyle, J. (2004). Effects of word frequency on recall memory following lorazepam, alcohol, and lorazepam alcohol interaction in healthy volunteers. *Psychopharmacology*, *176*(3-4), 420–425.

Stricker, J. L., Brown, G. G., Wixted, J., Baldo, J. V., & Delis, D. C. (2002). New semantic and serial clustering indices for the California Verbal Learning Test-Second Edition: background, rationale, and formulae. *Journal of the International Neuropsychological*

Society, *8*, 425–435.

Thalmann, B., & Monsch, A. U. (1997). CERAD. The consortium to establish a registry for Alzheimer’s disease. Neuropsychologische Testbatterie. Basel: Memory Clinic Basel.

Wilson, R. S., Bacon, L. D., Fox, J. H., Kramer, R. L., & Kaszniak, A. W. (2008). Word frequency effect and recognition memory in dementia of the alzheimer type. *Journal of Clinical Neuropsychology*, *5*, 97–104.

Zakzanis, K. K., Leach, L., & Kaplan, E. (1999). Dementia of the Alzheimer's type. In Kaplan E. F. & Leach, L. (Hrsg.) *Neuropsychological Differential Diagnosis* (pp. 33–33). Hove: Psychology Press.

TABELLE 1. Berechnung und Interpretation der Hauptvariablen des Auditiven Wortlisten Lerntests (AWLT).

Lernsumme
Stellt die Anzahl der insgesamt unmittelbar richtig widergegebenen Wörter in den vier Durchgängen des ersten Subtests dar. Die richtigen Antworten der vier Durchgänge werden dafür summiert.

Höhere Werte deuten auf eine intakte Enkodierungs- und Abrufleistung hin. Niedrige Werte indizieren jedoch eine genauere Inspektion anderer Variablen wie verzögerter Abruf und Wiedererkennen, um die zugrundeliegenden gestörten Prozesse zu identifizieren.

Kurz verzögerter Abruf
Bezeichnet die Anzahl der richtig erinnerten Wörter der gelernten Wortliste im zweiten Subtest nach einer Pause von ca. fünf Minuten.

Bei Gesunden Personen ist ein leichtes Nachlassen im Vergleich zum vierten Lerndurchgang möglich. Niedrige Werte gemeinsam mit einer verringerten Lernsumme deuten auf eine Störung von Aufmerksamkeit, Lernen, Enkodieren und/oder Speichern hin. Bei gleichzeitig hoher Lernsumme liegt eine Abrufstörung vor.

Lang verzögerter Abruf
Die Summe der richtig erinnerten Wörter der Wortliste im dritten Subtest nach einer Pause von ca. 20 Minuten.

Die Werte in dieser Variable liegen üblicherweise etwas unter denen von Kurz verzögertem Abruf und Lerndurchgang 4. Bei niedrigen Werten muss wie beim Kurz verzögerten Abruf die Interpretation im Kontext der Lernsumme geschehen werden, um Aufschluss über das zu Grunde liegende Defizit zu erhalten.

Wiedererkennen
Die Variable „Wiedererkennen" ist ein Maß zur Signal-Entdeckungstheorie, das die Fähigkeit beschreibt zwischen Zielwörtern und Distraktoren zu unterscheiden. Die Variable setzt sich aus der Anzahl der richtig eingeschlossenen Wörtern (WRE; Richtig-Positiv) und der Anzahl der falsch eingeschlossenen Wörter (WFE; Falsch-Positiv) im vierten Subtest über die folgende Formel zusammen:

$$\mathbf{WDI} = \mathbf{ln}\left(\frac{\left(\frac{\mathbf{WRE}+\mathbf{0,5}}{\mathbf{13}}\right) * \left(\mathbf{1}-\frac{\mathbf{WFE}+\mathbf{0,5}}{\mathbf{13}}\right)}{\left(\mathbf{1}-\frac{\mathbf{WRE}+\mathbf{0,5}}{\mathbf{13}}\right) * \left(\frac{\mathbf{WFE}+\mathbf{0,5}}{\mathbf{13}}\right)}\right)$$

Niedrige Werte resultieren aus Enkodierungs- und/oder Speicherstörungen und resultieren aus einem Response Bias mit vielen Hits (Richtig Eingeschlossen) und vielen Falsch-Positiven (Falsch Eingeschlossen) oder wenigen Hits (Richtig Eingeschlossen) in Kombination mit vielen Falsch-Positiven (Falsch Eingeschlossen). Daher sollte dieser Kennwert im Kontext der „Antworttendenz Ja" und der Hits (Richtig Eingeschlossen) interpretiert werden.

Index Verbales Langzeitgedächtnis
Um die Konstruktvalidät des Test-Sets Kognitive Funktionen Demenz (CFD) zu überprüfen, wurde ein Strukturgleichungsmodell mit den Hauptvariablen aller Tests gerechnet. Die Hauptvariablen des AWLT luden alle und als einzige auf den Faktor Verbales Langzeitgedächtnis. Der Indexwert ist der Summenwert der vier Hauptvariablen, die jeweils gewichtet nach ihrer Ladung auf den Faktor im Strukturgleichungsmodell eingehen und stellt

somit eine Gesamteinschätzung verschiedener Aspekte des verbalen Langzeitgedächtnisses dar.

Anmerkung. Texte teilweise adaptiert aus dem Manual des AWLT (Hessler & Jahn, 2017) und der Handanweisung des CFD (Jahn & Hessler, 2017).

TABELLE 2. Ergebnisse der RM-ANOVAs für den Einfluss von Wortlänge, Worthäufigkeit und orthografische Nachbarschaft auf die Abrufraten in den Lerndurchgängen 1 und 4 sowie lang verzögertem Abruf für Probanden mit und ohne Demenz vom Alzheimer-Typ.

Effekte	*F* (*df*, df_{Fehler})	*p*	Partial η^2
Wortlänge			
Wortlänge	**4.14 (1, 48)**	**0.048**	**0.08**
Wortlänge × Diagnose	1.22 (1, 48)	0.275	0.03
Subtest	77.73 (2, 96)	< 0.001	0.62
Subtest × Wortlänge	**10.85 (2, 96)**	**< 0.001**	**0.18**
Subtest × Diagnose	20.79 (2, 96)	< 0.001	0.30
Subtest × Wortlänge × Diagnose	0.78 (2, 96)	0.461	0.02
Diagnose	35.11 (1, 48)	< 0.001	0.42
Worthäufigkeit			
Worthäufigkeit	**16.36 (1, 48)**	**< 0.001**	**0.25**
Worthäufigkeit × Diagnose	0.58 (1, 48)	0.450	0.12
Subtest	69.88 (2, 96)	< 0.001	0.59
Subtest × Worthäufigkeit	**11.25 (2, 96)**	**< 0.001**	0.19
Subtest × Diagnose	22.32 (2, 96)	< 0.001	0.32
Subtest × Worthäufigkeit × Diagnose	2.74 (2, 96)	0.069	0.05
Diagnose	36.14 (1, 48)	< 0.001	0.43
Orthografische Nachbarschaft			
Nachbarschaft	**13.06 (1, 48)**	**0.001**	**0.21**
Nachbarschaft × Diagnose	3.93 (1, 48)	0.053	0.08
Subtest	69.05 (2, 96)	< 0.001	0.59
Subtest × Nachbarschaft	0.41 (2, 96)	0.662	0.01
Subtest × Diagnose	15.98 (2, 96)	< 0.001	0.25
Subtest × Nachbarschaft × Diagnose	1.89 (2, 96)	0.157	0.04
Diagnose	31.58 (1, 48)	< 0.001	0.40

Anmerkung. Statistisch signifikante Haupt- und Interaktionseffekte mit linguistischen Variablen sind **fettgedruckt**. *df* = Freiheitsgrade

TABELLE 3. Post-hoc *t*-tests für die signifikanten Interaktionseffekte der RM-ANOVAs.

Effekt	*M* (*SD*)	Differenz *M* (*SE*)	Differenz 95% CI	*t* (*df*), *p*	Cohens *d*
Subtest × Wortlänge					
Lerndurchgang 1					
Kurz	16.00 (24.50)	−14.89 (19.90)	−20.54; −9.23	−5.29 (49), < 0.001	0.73
Lang	30.89 (15.44)				
Lerndurchgang 4					
Kurz	47.33 (30.18)	−7.34 (28.54)	−15.44; 0.78	−1.82 (49), 0.075	0.24
Lang	54.67 (22.76)				
Lang verzögerter Abruf					
Kurz	38.00 (36.27)	6.67 (29.44)	−1.70; 15.03	1.60 (49), 0.116	−0.20
Lang	31.33 (29.50)				
Subtest × Worthäufigkeit					
Lerndurchgang 1					
Selten	13.33 (21.30)	−18.45 (24.86)	−25.51; −11.38	−5.25 (49), < 0.001	0.92
Häufig	31.78 (18.51)				
Lerndurchgang 4					
Selten	43.33 (30.30)	−12.67 (24.69)	−19.66; −5.68	−3.64 (59), 0.001	0.48
Häufig	56.00 (21.53)				
Lang verzögerter Abruf					
Selten	33.33 (35.63)	0.44 (24.28)	−6.45; 7.34	0.13 (49), 0.898	−0.01
Häufig	32.89 (28.57)				

Anmerkung. *M* = Mittelwert, *SD* = Standardabweichung, *SE* = Standardfehler, 95 % CI = 95 % Konfidenzintervall, *df* = Freiheitsgrade.

TABELLE 4. Differentielle Validität des AWLT für die Unterscheidung zwischen Unbeeinträchtigten und Patienten mit Demenz vom Alzheimer-Typ.

Kennwert	Cohens *d*	AUC (95% CI)
Lernsumme	−1.55	0.87 (0.76; 0.97)
Kurz verzögerter Abruf	−2.29	0.94 (0.88; 1.00)
Lang verzögerter Abruf	−2.02	0.92 (0.84; 1.00)
Wiedererkennen	−1.87	0.89 (0.81; 0.98)
Index Verbales Langzeitgedächtnis	−2.17	0.94 (0.88; 1.00)

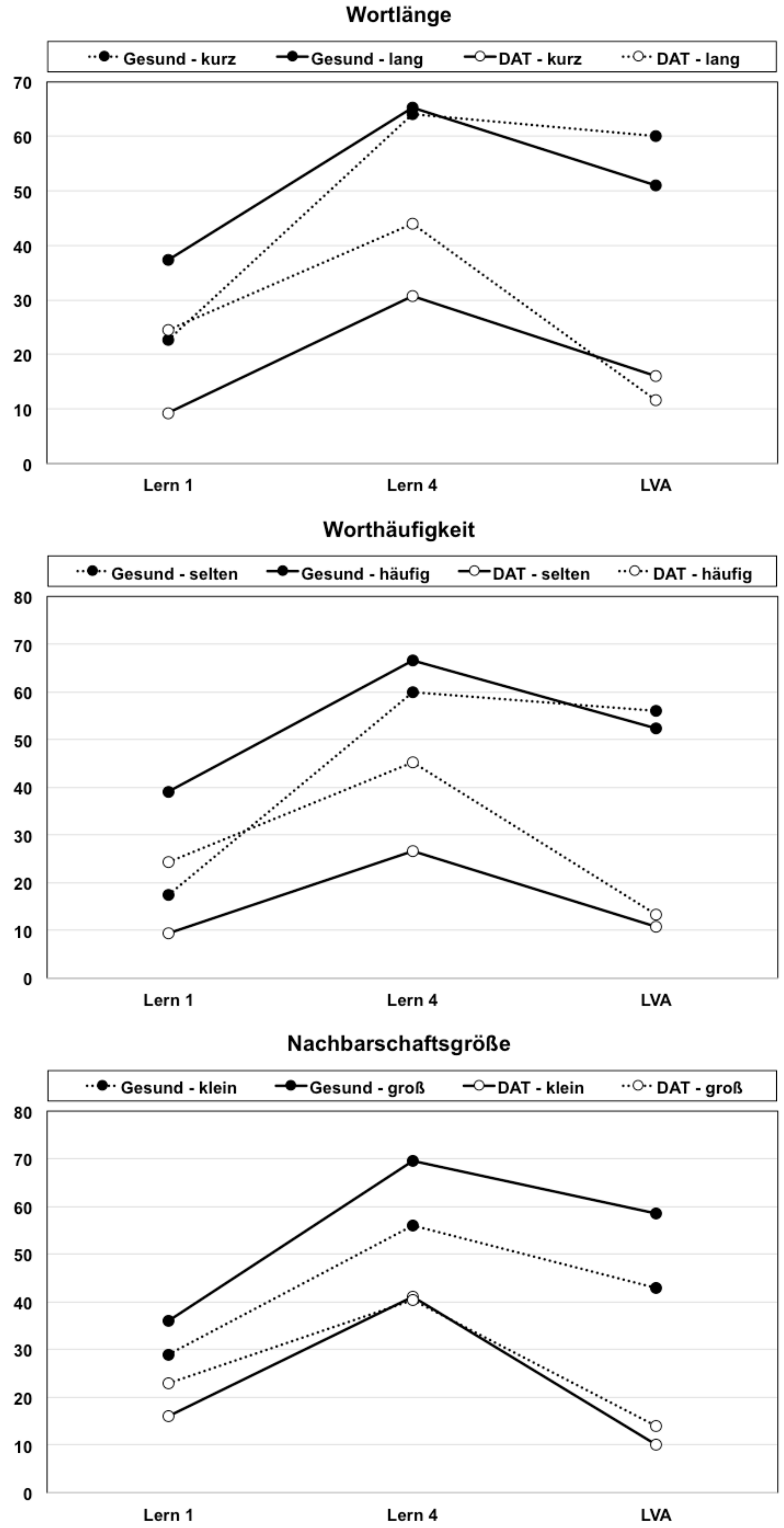
Anmerkung. AUC = area under the curve, 95 % CI = 95 % Konfidenzintervall.

TABELLE 5. Vergleich der linguistischen Gedächtniseffekte im deutschen California Verbal Learning Test (CVLT) und dem Auditiven Wortlisten Lerntest (AWLT). Pfeile nach oben bedeuten einen Vorteil des AWLT gegenüber dem CVLT.

Linguistische Variable	Test	Effekt in RM-ANOVA	Partial η^2	Bewertung AWLT
Wortlänge	CVLT	-	-	
	AWLT	Haupteffekt	0.08	↓
		Interaktion mit Subtest	0.30	
Häufigkeit	CVLT	Haupteffekt	0.70	↑
		Interaktion mit Subtest	0.05	
	AWLT	Haupteffekt	0.25	
		Interaktion mit Subtest	0.19	
Nachbarschaftsgröße	CVLT	Interaktion mit Diagnose	0.70	↑
	AWLT	Haupteffekt	0.21	

Anmerkung. RM-ANOVA = Varianzanalyse mit Messwiederholungen.

ABBILDUNG 1. Abrufraten (%) der Wörter des AWLT in Abhängigkeit von Diagnose, Subtest und linguistischer Eigenschaft.



Anmerkung. DAT = Demenz vom Alzheimer-Typ, Lern 1 = Lerndurchgang 1, Lern 4 = Lerndurchgang 4, LVA = lang verzögerter Abruf.

TABELLE I. Ablauf des Auditiven Wortlisten Lerntests (AWLT).

Subtest	Dauer in Minuten
Lernphase Durchgänge 1 – 4; auditive Präsentation	7
Pause Im CFD ersetzt mit WAFA. Bei alleinstehender Durchführung soll hier eine Pause eingelegt oder diese mit nonverbalen, nicht gedächtnisbezogenen Tests ersetzt werden.	5
Kurz verzögerter Abruf	2
Pause Im CFD ersetzt mit WAFG, TMT und CORSI. Bei alleinstehender Durchführung soll hier eine Pause eingelegt oder diese mit nonverbalen, nicht gedächtnisbezogenen Tests ersetzt werden.	20
Lang verzögerter Abruf	2
Wiedererkennen Auditive Präsentation der 12 Wörter unter 12 linguistisch und semantisch parallelisierten Distraktoren.	2
Gesamtdauer	38 – 40

Anmerkung. CFD = Test-Set Kognitive Funktionen Demenz. WAFA = Wahrnehmungs- und Aufmerksamkeitsfunktionen - Alertness, WAFG Wahrnehmungs- und Aufmerksamkeitsfunktionen – Geteilte Aufmerksamkeit, TMT = Trail Making Test, CORSI = Corsi-Block-Tapping-Test.