Article

SAGE Open
October-December 2016: 1–6

DOI: 10.1177/2158244016681825
sgo.sagepub.com
SAGE

Prevalence of Two-Syllable Digits Affecting Forward Digit Span Test Score: A Potential Reliability Factor in Digit Span Tests and New Light to the Word Length Effect

Lars E. Egner[1], Stefan Sütterlin[1,2], and Ricardo G. Lugo[1]

Abstract

The word length effect shows a connection between word length and working memory performance. Although the relationship between digit verbal length and digit span has been investigated between languages, it has not been investigated within a language. It was hypothesized that this effect can be shown as a connection between the prevalence of digits with two syllables and digit span score. The study examined the effect of amount of syllables on Norwegian digit span test scores by altering the prevalence of two-syllable digits using three conditions in a repeated measures design (N = 54). Results suggest that an elimination of two-syllable digits in a digit span test significantly reduced forward span test score (Cohen's d = 0.36), but had no effect on backward span scores. These results suggest that a balanced distribution of two-syllable digits in a forward digit span tests should theoretically increase the test's comparability and reduce language-related biases thus increasing the test's parallel-form reliability. A peak-span model is proposed to integrate the findings into previous research on the word length effect.

Keywords

digit span, two-syllable digits, word length effect, working memory, peak-span model

Introduction

Assessing working memory's capacity is of high relevance in psychology. Forward digit span (FDS) and backward digit span (BDS) are widely used measures of working memory (Richardson, 2007). The easy administration of both the FDS and BDS, instant scoring, and minimal need for equipment make the test a popular choice in both clinical testing and empirical research and is a part of the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1997). The WAIS is used to evaluate individuals in multiple settings, ranging from job-related assessment centers, criminal responsibility to organizational membership. The test is also used in research to demonstrate individual cognitive differences, most recently in the emerging field of environmental psychology (Berman, Jonides, & Kaplan, 2008; Kaplan & Kaplan, 1989).

The digit span (DS) test, though widely used, suffers from some methodological limitations. First, the test is presented orally, and thus clarity, pitch, and rhythm in pronunciation could affect the scores (Silverman, 2007). Second, the scoring is rather problematic. Woods and colleagues (2010) explored more than three common methods of scoring and testing, with the most controversial one ("two-error total trials") achieving a test–retest reliability correlation as low as r_{tt} = .12. The current WAIS-IV FDS test has a test–retest reliability of r_{tt} = .74 (Wechsler, 2008), which is considered acceptable.

The word length effect describes a negative correlation between the pronunciation length of a word, often measured in syllables, and one's ability to recall the word over a short period of time. For example: the words "university," "tuberculosis," and "vegetarian" are harder to remember than "pen," "ice," and "leaf" (Baddeley, Thomson, & Buchanan, 1975). This has been postulated to be due to that the phonological loop has a limited capacity and therefore phonetically longer words are harder to store (Todd &

[1]Lillehammer University College, Norway
[2]Oslo University Hospital, Rikshospitalet, Norway

Corresponding Author:
Lars E. Egner, Lillehammer University College, P.O. Box 952, Lillehammer N-2604, Norway.
Email: Lars.egner91@gmail.com

Marois, 2004). The underlying mechanisms and exact boundaries for this phenomenon have been topics for recent debates (see Hulme, Suprenant, Bireta, Stuart, & Neath, 2004; Jalbert, Neath, Bireta, & Surprenant, 2011). The amount of phonological information that can be stored has been reported to be roughly equivalent to how much a person can articulate in 2 s (Baddeley, 2012).

Despite debates discussing the length of words and digits' influence on phonological processing, the brevity of digits used in DS tests has never been statistically controlled. Phonetic effects on DS have been studied when comparing languages, that is, Welsh versus English (Ellis & Hennelly, 1980) or Chinese versus non-Chinese speakers (Tang et al., 2006), but not within one language.

These interlanguage differences may be diminished over many tests but this may compound errors created within individual tests. The law of large numbers will distribute a large amount of digits evenly, but a single DS test may deviate from this. This could lead to single tests often being easier or harder than others decreasing the reliability of parallel forms (Kaplan & Saccuzzo, 2013). A test will be perfectly comparable with tests with identical numerical compositions, but not other versions containing other combinations. Digit span tests that are generated randomly for each participant, will on average have a valid score, but the variability of individual scores as well as the participants required to achieve statistically valid results will suffer.

Due to the word length effect and the varying prevalence of two-syllable digits (TSD) present in a DS test, the authors suggest that DS tests could have a varying difficulty related to the prevalence of TSD in the DS test. This study will explore the prevalence of TSD effect on DS score within a language.

Method

Design

The Norwegian language contains two digits out of 10 that have two syllables ("fi|re" = four; "åt|te" = eight). The digit zero is not part of most DS tests and was excluded. The digits four and eight were used as TSDs.

The study followed a single-blind, counterbalanced experimental repeated measures design featuring two experimental conditions and one control condition. The first condition, the no-TSD condition, had no prevalence of TSD in the tests given. The second condition, the normal-TSD condition, had two of nine digits with TSD present in the tests. The third condition, the double-TSD condition, had a double prevalence (four of 11) of TSD present. The normal-TSD condition was treated as the control condition.

The first session included three handmade DS tests. A second session was conducted using a computer generated algorithm to create individual tests for each participant. This was to avoid the systematic bias of having one set of tests so that the only difference between conditions was the prevalence of TSDs. The results from the two sessions showed no statistically significant differences in all conditions, no-TSD, $t(52) = 0.43$, $p = .67$, $d = 0.11$; normal-TSD, $t(52) = 0.04$, $p = .97$, $d = 0.01$; double-TSD, $t(52) = 0.22$, $p = .84$, $d = 0.06$. Data sets were merged since there were no statistical significant differences.

Participants

University students ($N = 54$; female = 30) were recruited to participate (mean age = 22.6; $SD = 2.8$). Participants were excluded from the study if (a) they had been consciously practicing memorizing digits for more than 2 hr in their life, (b) were not native speakers, or (c) were familiar with the experiment's purpose.

Procedure

The procedure for both the forward and backward span testing was based on the WAIS-III (Wechsler, 1997). The participants were informed that they are free to leave at any time without providing any reasons and ensured of the anonymization of all data in accordance to the institution's ethical requirements. The participants received the verbal instructions according to WAIS-III: "I will now say some digits, listen closely and repeat them back to me when I am done." An example was given, "so if I say 1-2-3, you will say . . . ?" and the test started with the experimenter verbally reading the digits to the participant. Testing took 10 to 20 min depending on performance. Participants were debriefed after testing and were asked not to disclose the purpose of the experiment to other participants. Comments or feedback from the participants was also recorded and coded.

Materials

Tests for the first session were primarily created by shuffling the digits in random order and then rearranging them if they broke several rules. Identical digits could not appear next to each other, neighboring digits could not be within one digit of each other (e.g., 1-2-5), digit strings could not have a ascending or descending order (e.g., 2-4-6), digits could not be repeated within three digits (e.g., 7-3-7), and finally digit strings could not start or end with the same digits of any adjacent strings. The algorithm followed all of the rules mentioned, but because the algorithm distributed the digits very equally, it was a common occurrence that certain sequences repeated (e.g., 8-3-6 . . . 8-3-6), which made the test easier. This resulted in a rule that was included in the algorithm stated that when two digits had occurred next to each other, they could not appear in the same order for 15 digits (e.g., 2-6 . . . 2-6). The prevalence of each digit could easily be manipulated in the program, and each participant had three unique DS tests, ensuring that the only stable difference among the

tests was the prevalence of TSDs. The algorithm used to create the tests can be downloaded at www.researchgate.net/publication/303985006_Digit-Span_Number_Generator.

Tests were scored by giving one point per successful digit string, in line with the WAIS-III administration manual (Wechsler, 1997). This scoring method is the most common, and would be generalizable to most research and testing settings. If a participant could constantly repeat up to seven digits but not any seven-digit strings, their score would be 10, one point for each previous string. Strings with one digit are not a part of the test.

Statistical Analysis

All data were treated as continuous data and processes in IBM's SPSS 20 (IBM Corp, 2011). A one-way repeated measures ANOVA was used to investigate the effect of the independent variable prevalence of TSD on the dependable variable DS score with three conditions, no TSD, normal amount of TSD, and double amount of TSD.

A least significant difference (LSD) test with Bonferroni correction with alpha levels of .017 (.05/3) was used for post hoc test.

Scores were checked for normality and checked for outliers using the outlier labeling rule, multiplying the interquartile range by a factor of 2.2 (Hoaglin, Iglewicz, & Tukey, 1986). No outliers were found.

Results

Means and standard error are presented in Figure 1. There was a significant main effect of prevalence of TSD on FDS score Wilks's Lambda = .806, $F(2,52) = 6.27$, $p = .004$, partial $\eta^2 = .194$. LSD tests indicated that the average score was significantly lower in the no-TSD condition ($M = 8.22$, $SD = 1.49$) than were those in both the normal-TSD ($M = 8.76$, $SD = 1.52$), $t(53) = 2.91$, $p = .005$, $d = 0.36$, and double-TSD ($M = 8.72$, $SD = 1.62$), $t(53) = 2.53$, $p = .014$, $d = 0.32$, conditions. Normal-TSD and double-TSD conditions did not differ in test scores, $t(53) = 0.15$, $p = .88$, $d = 0.03$.

Contrary to the significant results regarding FDS and TSD, the same was not found for the BDS test and TSD, Wilks's Lambda = .987, $F(2, 28) = .182$, $p = .835$, $\eta^2 = .013$; no-TSD condition scored $M = 5.37$ ($SD = 1.77$), normal-TSD $M = 5.43$ ($SD = 1.91$), double-TSD $M = 5.56$ ($SD = 2.18$). The prevalence of TSD had no influence on BDS.

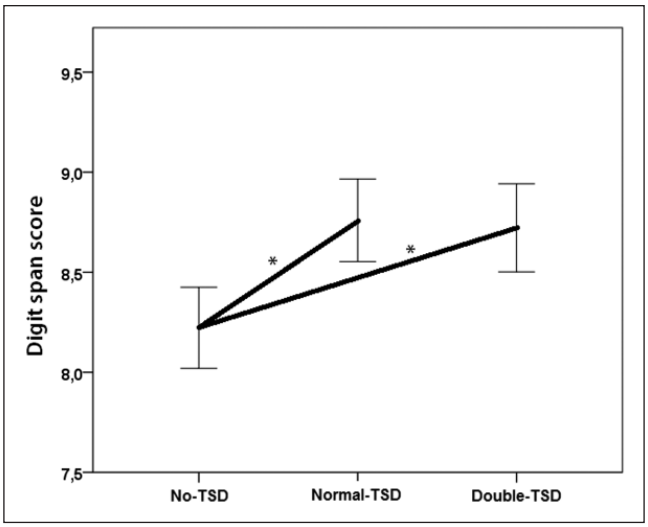

Figure 1. FDS score showed significant differences between the no-TSD, normal-TSD, and double-TSD condition.
Note. Bars represent standard error. FDS = forward digit span; TSD = two-syllable digits.
*$p < .05$.

Discussion

Cognitive Explanation

The results indicate a positive association between digit brevity and working memory performance, contradicting earlier findings on the word length effect (Baddeley et al., 1975). However, the results do not necessarily contradict previous findings on the word length effect as the manipulation of the word or digit brevity differs from previously conducted research. The authors argue that these findings are complementary to concurrent word length effect research by adding findings suggested by the Level of Processing Theory formulated by Craik and Lockhart (1972).

A different phonological loop functionality perspective. Word length experiments suggest a relationship between the amount of syllables in a word and the probability of recall, suggesting a linear association between load on the phonological loop and decay (Baddeley, 2012). We argue that the phonological loop can handle a certain amount of load *before* memory content decays. Although the phonological loop displays the same amount of decay at two and five syllables, there is less decay on five syllables than in 30 syllables. The phonological loop's capacity might thus be sufficient to process a number of syllables up to a certain limit and, only after this point does information start to decay, dubbed *phonological loop overload*. This peak of best performance probably varies individually.

Previous studies were not designed to detect or investigate this effect, because classical word length effect experiments compare five words with five syllables to five words with 10 syllables, missing out on potential differences in performance on different combinations (e.g., five words with six syllables to five words with seven syllables etc.). These more subtle differences are explored with the manipulated material in the present study, suggesting increased performance with more syllables, at least up to a certain peak. Although this effect can explain the maintained performance between the conditions with only single syllable digits

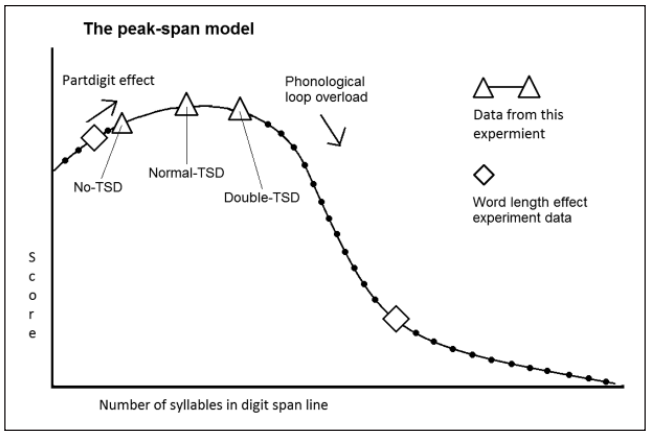

Figure 2. The peak-span model, an interaction of the part-digit effect and Phonological Loop Overload, a possible explanation of the findings.
Note. Diamond shapes represent data points from classical word length effect experiments. TSD = two-syllable digits.

compared with the standard or double amount, the *part-digit effect* can account for the performance *increase* from no-TSD to normal-TSD and double-TSD.

The suggested part-digit effect proposes that a full recall of TSD can be triggered by one remaining syllable even after loss of a second syllable in situations where the initial syllable is unique in the amount of possible words one syllable can be used to recall the whole digit (here: the digits 1 to 9). The part-digit effect could rely on the same mechanisms of retrieval cue research (Tulving & Osler, 1968) where pairing words to other words facilitated recall of the original word.

The peak-span model (PSM). The suggested part-digit effect and phonological loop overload could account for the phenomenon the present study shows as well as word length effects shown in previous studies (e.g., Baddeley et al., 1975). In the no-TSD condition, the absence of the part-digit effect lowers the score. In the normal-TSD condition, the part-digit effect gives an increase in score, but the phonological loop has not yet reached a point where phonological loop overload significantly reduces score. In the double-TSD condition, the part-digit effect further increased scores, but the increased load on the phonological loop (word length effect) caused a greater loss of information resulting in an equal score as the normal-TSD condition. These combined effects are referred to as peak-span; a model is shown in Figure 2.

The two proposed mechanics could account for the findings presented in this study. The part-digit effect could increase the recall of digits since they are longer and digit brevity does not cause phonological loop overload at very low loads, such as five syllables. The PSM is not exclusive to other models of immediate serial recall, such as the Primacy Model (Page & Norris, 1998), as its proposed mechanisms are only relevant in a specific low mix of digits and syllables.

As the amount of syllables had no effect on the backward span, this could indicate that the BDS test is not as dependent on the phonological loop as the FDS test, but to a larger degree involves the central executive component of working memory as BDS requires manipulation of items (Giofrè, Stoppa, Ferioli, Pezzuti, & Cornoldi, 2016).

Improving the DS Test

Results suggest a positive relationship between the prevalence of syllables in a FDS test and FDS score. Although these preliminary results and the design of the study do not allow for the assumption of a linear correlation, the findings indicate that a lower than standard amount (22.2% in Norwegian) of TSD can affect DS test scores negatively. The results suggest that statistically controlling for the prevalence of TSD in FDS could theoretically improve the comparability of different language forms of this very frequently applied test. DS tests with higher frequencies of TSD could rely on different cognitive processes that may lead to score differences than lower TSD frequency tests. This could decrease the tests overall validity as a result of lacking comparability, and thus impairing its parallel-form reliability. Parallel-form reliability is defined as comparing two equivalent forms of a test that measure the same attribute (Kaplan & Saccuzzo, 2013). Confirming the parallel-form reliability improvement can only be done by a separate study investigating this. To minimize biases, TSD need to be equally distributed throughout the test.

Although an effect size of Cohen's $d = 0.36$ is small in most studies using DS tests as a dependent variable, the authors argue that effect sizes must be interpreted in the correct context (Thompson, 2009). In this study's context, this effect size is not representing an external factor's influence on FDS, but rather intra-individual variability within the same test, which ideally should be close to zero. Explaining 19.4% of the variance is especially relevant given the large scale on which DS tests are used in epidemiological studies and as part of routine assessment in numerous contexts and with large samples.

The conclusions based on these data have to be interpreted with care, given the moderate sample size. Further research is needed to replicate these findings in other languages with different amounts of TSD or even more syllables per digit.

Two one-way ANOVAs were employed instead of a two-way ANOVA because the second sample did not receive the BDS test. Incidentally an interaction effect has not been investigated.

It should not be ruled out that the same results could be achieved by altering the prevalence of random digits. It is possible that the scores were lower in the no-TSD condition because participants were implicitly expecting the digits four and eight to appear. In the no-TSD condition, all TSDs were removed, but it could be that removing two one-syllable

digits would have the same effect. One could argue that if valid, this expectancy issue should also have influenced the BDS scores, but these showed no difference. Although it is likely that FDS and BDS rely on different stores, the expectancy issue could be valid for FDS and not for BDS. This possibility should be investigated before committing to further TSD/FDS research.

Further research could attempt to apply this model by using the auditory length (actual length) of digit words, rather than the amount of syllables as their commonly used correlate.

Because the DS test is included in several test batteries (e.g., Wechsler Intelligence Scale for Children [WISC], WAIS), it is possible that an increase in the instrument's parallel-form reliability could increase the loading of the test to the general intelligence factor *g*. Future research is needed to investigate the consequences of revised DS assessment on *g*-loading.

It would also be interesting to investigate aspects related to the word length effect. Assumption of a linear association between word length and DS performance challenges common interpretations of word length effect theory. The proposed phonological loop overload mechanisms suggest that any number of syllables under a peak is equally hard to remember, but anything beyond this number is subject to decay. For example, the word string "Tree, grass, cup, lake," would be equally hard or harder to remember than the word string "Tree, grass, cup, river."

Conclusion

Theory surrounding word length effect should be extended to explain the presented data, a curvilinear relationship between digits oral length and FDS score. The suggested phonological loop overload mechanism and part-digit effect suggest that the phonological loop may handle smaller loads of information at a maximum efficiency, and that one syllable of a digit can be used to recall a full digit. These effects together form the PSM, suggesting a reverse-U curvilinear relationship between syllables in a DS string and DS score. Investigating whether aspects of the PSM are relevant for words could also be of interest.

The study shows that there is sufficient evidence to suggest that using a strict statistical distribution of TSD when generating FDS tests could improve the test. The reported findings should be replicated also in other languages other than Norwegian and other measures of word length (auditory length) could be applied. We further suggest to investigate word length effect theory under consideration of equal processing of any number of syllables up to an individual peak.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1-29. doi:10.1146/annurev-psych-120710-100422

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. doi:10.1016/S0022-5371(75)80045-4

Berman, M. G., Jonides, J., & Kaplan, S. (2008). The cognitive benefits of interacting with nature. *Psychological Science*, *19*, 1207-1212. doi:10.1111/j.1467-9280.2008.02225.x

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671-684. doi:10.1016/S0022-5371(72)80001-X

Ellis, N. C., & Hennelly, R. A. (1980). A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology*, *71*, 43-51. doi:10.1111/j.2044-8295.1980.tb02728.x

Giofrè, D., Stoppa, E., Ferioli, P., Pezzuti, L., & Cornoldi, C. (2016). Forward and backward digit span difficulties in children with specific learning disorder. *Journal of Clinical and Experimental Neuropsychology*, *38*, 478-486. doi:10.1080/13803395.2015.1125454

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *81*, 991-999. doi:10.1080/01621459.1986.10478363

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. doi:10.1037/0278-7393.30.1.98

IBM Corp. (2011). IBM SPSS Statistics for Windows, Version 20.0 [Released]. Armonk, NY: Author.

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. doi:10.1037/a0021804

Kaplan, R. M., & Kaplan, S. (1989). *The experience of nature: A psychological perspective*. Cambridge, UK: Cambridge University Press.

Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological testing: Principles, applications, and issues* (8th ed.). Belmont, CA: Wadsworth.

Page, M., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, *105*, 761-781. doi:10.1037/0033-295X.105.4.761-781

Richardson, J. T. E. (2007). Measures of short-term memory: A historical review. *Cortex*, *43*, 635-650. doi:10.1016/S0010-9452(08)70493-3

Silverman, M. J. (2007). The effect of paired pitch, rhythm, and speech on working memory as measured by sequential digit recall. *Journal of Music Therapy*, *44*, 415-427. doi:10.1093/jmt/44.4.415

Tang, Y., Zhang, W., Chen, K., Feng, S., Ji, Y., Shen, J., …Liu, Y. (2006). Arithmetic processing in the brain shaped by cultures. *Proceedings of the National Academy of Sciences*, *103*, 10775-10780. doi:10.1073/pnas.0604416103

Thompson, B. (2009). A brief primer on effect sizes. *Journal of Teaching in Physical Education*, *28*, 251-254.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *428*, 751-754. doi:10.1038/nature02466

Tulving, E., & Osler, S. (1968). Effectiveness of retrieval cues in memory for words. *Journal of Experimental Psychology*, *77*, 593-601. doi:10.1037/h0026069

Wechsler, D. (1997). *WAIS-III: Administration and scoring manual* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale: WAIS-IV; Technical and Interpretive Manual*. San Antonio, TX: Pearson.

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., . . . Reed, B. (2010). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, *33*, 1-11. doi:10.1080/13803395.2010.493149

Author Biographies

Lars E. Egner researches within Lillehammer University College's Masters program in Environmental Psychology. His research interests encompass cognitive processes related to exposure to nature and the role of environmental influences for perception and restoration.

Stefan Sütterlin is professor in psychology at Lillehammer University College and Oslo University Hospital. His research interests include inhibitory processes, cognitive performance, and emotion regulation.

Ricardo G. Lugo is health psychologist and head of psychology department at Lillehammer University College. He also works as consultant for professional athletes and researches amongst others on pain, physical activity, and cognitive performance.