Is the Hebb repetition task a reliable measure of individual differences in sequence learning?

Louisa Bogaerts[1,2,3], Noam Siegelman[3], Tali Ben-Porat[3],
& Ram Frost[3,4,5]

[1]Ghent University, Ghent, Belgium
[2]CNRS & Aix-Marseille University
[3]Hebrew University, Jerusalem, Israel
[4] Haskins Laboratories, New Haven, USA
[5] The Basque Center for Brain and Language, San Sebastian, Spain

Corresponding author:

Dr. Louisa Bogaerts
Department of Psychology, The Hebrew University, Jerusalem 91905, Israel
E-mail: bog.louisa@gmail.com

Word count: 5739

Acknowledgement

This work was supported by the Israel Science Foundation (Grant 217/14 awarded to Ram Frost), the National Institute of Child Health and Human Development (Grant RO1-HD 067364 awarded to Ken Pugh and Ram Frost, PO1-HD 01994 awarded to Haskins Laboratories), ERC-2015-AdG-692502, and by the Research Foundation-Flanders/The Fyssen foundation, of which Louisa Bogaerts was a research fellow. We thank Wouter Duyck and Arnaud Szmalec for their contribution to and helpfull comments on earlier versions of this paper.

Abstract

The Hebb repetition task, an operationalization of long-term sequence learning through repetition, is the focus of renewed interest, as it is taken to provide a laboratory analogue for naturalistic vocabulary acquisition. Indeed, recent studies have consistently related performance in the Hebb repetition task with a range of linguistic (dis)abilities. However, in spite of the growing interest in the Hebb repetition effect as a theoretical construct, no previous research has ever tested whether the task used to assess Hebb learning offers a stable and reliable measure of individual performance in sequence learning. Since reliability is a necessary condition to predictive validity, in the present work we tested whether individual ability in visual verbal Hebb repetition learning displays basic test-retest reliability. In a first experiment Hebrew-English bilinguals performed two verbal Hebb tasks, one with English and one with Hebrew consonant letters. They were retested on the same Hebb tasks after a period of about six months. Overall serial recall performance proved to be a stable and reliable capacity of an individual. By contrast, the test-retest reliability of individual learning performance in our Hebb task was close to zero. A second experiment with French speakers replicated these results and demonstrated that the concurrent learning of two repeated Hebb sequences within the same task minimally improves the reliability scores. Taken together, our results raise concerns regarding the usefulness of at least some current Hebb learning tasks in predicting linguistic (dis)abilities. The theoretical implications are discussed.

Keywords: sequence learning, serial recall, the Hebb repetition effect, individual differences, test reliability

Introduction

In the early 1960s, Donald Hebb (1961) asked his participants to perform an immediate serial recall task in which one specific sequence of digits was repeated every third trial (unannounced). In his influential paper, Hebb reported that, over a number of trials, participants' recall performance for the repeated sequence improved relative to the nonrepeating sequences. This effect was later labeled as the 'Hebb Repetition Effect' (HRE). In essence, the HRE reflects how a sequence of information in short-term memory gradually develops into a more stable, long-term memory trace, through repeated presentation and recall. The Hebb effect has been replicated in many studies involving young and older adults (e.g., Cumming, Page, & Norris, 2003; Turcotte, Gagnon, Poirier, 2005) as well as children (e.g., Gould & Glencross, 1990; Mosse & Jarrold, 2008; Smalle et al., 2015), across sensory modalities (*visual*: e.g., Page, Cumming, Norris, Hitch & McNeil, 2006; *auditory*: e.g., Parmentier, Maybery, Huitson, & Jones, 2008). The task variants used in the literature vary in their specific parameters such as the stimulus material (e.g., *letters*: e.g., Page et al., 2006; *syllables:* e.g., Szmalec, et al,, 2009; *words:* e.g., Sechler & Watkins, 1991; *spatial locations*: e.g., Couture & Tremblay, 2006), list length (typically ranging from 6 to 9 items) and/or presentation rate of stimuli, the method for repeating the Hebb sequence (e.g., *full repetition*: Page et al., 2006; *partial repetition*: Szmalec, et al., 2009), the response format (e.g., *verbal*: e.g., Mosse & Jarrold, 2008; *mouse clicking*: e.g., Page et al., 2006), thus exemplifying the wide context in which Hebb learning can be observed.

In the past two decades Hebb repetition learning was the subject of renewed interest. As an operational construct, the HRE was put forward as a laboratory analogue for the learning process involved in naturalistic vocabulary acquisition (e.g., Cumming et al., 2003; Page & Norris, 2009). In this view, new phonological word-forms are conceived as memorized sequences

of sublexical units (phonemes, syllables) through repeated exposure (Page & Norris, 2009; Szmalec et al., 2009). In support of this claim, recent work showed that presenting participants with printed syllabic sequences such as "*la-va-bu*" in the Hebb repetition paradigm, results in auditory lexical competition with existing words ("lavabo"), just like existing word-forms do (Szmalec, Page, & Duyck, 2012). This theoretical approach relates memory for serial order to language acquisition. In the context of reading acquisition and reading disorders for example, learning orthographical word-forms is taken to reflect the creation of long-term representations of repeated grapheme sequences through repeated exposure, and by extension, reading impairments would, therefore, be associated with a deficit in this long-term learning of serial-order information (Bogaerts, Szmalec, Hachmann, Page, & Duyck, 2015; Szmalec et al., 2011).

The empirical evidence supporting the theoretical link between serial-order memory and linguistic abilities hinges mainly on group studies. These studies typically focus on the average success rate of the sampled population in the Hebb paradigm, as measured by their increased success on the repeated trials relative to baseline performance on non-repeated fillers. In general, these studies have shown poorer serial-order learning abilities in a variety of clinical populations, such as adults with dyslexia, children with reading difficulties, or children with Specific Language Impairment (SLI), relative to matched samples of controls (Bogaerts et al., 2015; Gould & Glencross, 1990, Szmalec et al., 2011; Hsu & Bishop, 2014; but see Staels & Van den Broeck, 2014, for different results). In a similar vein, preservation of serial-order learning abilities as measured in the Hebb task was demonstrated in a sample of individuals with Down syndrome, who typically show relative strengths in vocabulary size (Mosse & Jarrold, 2010).

The interest in the HRE has further led to a series of correlational studies, aiming to examine whether individual differences in Hebb repetition performance could reliably predict

performance in language-related tasks. Most relevant is the observation of a positive correlation between individual Hebb learning performance and nonword learning, in a sample of typically developing children (Mosse & Jarrold, 2008, see also Archibald & Joanisse, 2012, for a similar finding). More recently, we also have reported (Bogaerts, Szmalec, De Maeyer, Page & Duyck, 2016) significant (albeit weak) positive correlations between the magnitude of Hebb repetition learning and reading performance in children. In contrast, Hsu and Bishop (2014) failed to find a significant correlation between individuals' Hebb learning performance and vocabulary scores or grammar abilities. Similarly, the Hebb repetition task has been used to study the learning of serial order information in some neurological patients. Gannon and colleagues, for example, assessed the Hebb learning ability of an amnesic patient and showed that his learning magnitude, as well as learning rate, were comparable to those observed in matched control participants (Gannon, Forster, Turcotte, & Jongenelis, 2004; see Jefferies, Bott, Ehsan & Lambon, 2011 for a similar single-case approach with a semantic dementia patient).

What all these recent correlational and single-case studies have in common is the underlying implicit assumption that participants' performance in the Hebb repetition task reflects a reliable and relatively stable individual capacity in memory for serial order. Moreover, the correlational studies assume that this ability should reliably predict a range of linguistic skills. Nevertheless, the observed correlations are often weak (e.g., Bogaerts et al., 2016) or even absent (Bishop and Hsu, 2014). Surprisingly however, in spite of the growing interest in serial order learning ability as measured through the Hebb repetition task, to our knowledge no research has ever tested whether individual abilities in serial-order learning operationalized by this task, are indeed stable and reliable measures (similar, for example, to measures of intelligence, working memory, or statistical learning performance). This question is not simply a methodological one

but has important theoretical implications. Test-retest reliability is a necessary condition for predictive validity. Any task that aims to predict a given cognitive function, must display test-retest reliability, for if not, participants' score in a given session may reflect either situation-specific or error variance (see Siegelman & Frost, 2015, for a similar discussion in the domain of statistical learning). It should be noted that in their paper reporting a positive correlation between Hebb learning performance and nonword learning, Mosse and Jarrold (2008) did look at the split-half reliability of Hebb repetition performance and reported a coefficient of 0.48. This type of reliability is important but concerns only the *internal consistency* of the measure and not its stability in time, or across testing materials.

The present study provides a first much-needed examination of the test-retest reliability of verbal Hebb repetition learning as an individual ability. It reports two experiments across two different populations, and two different experimental procedures. In *Experiment 1*, Hebrew-English bilinguals performed two verbal Hebb repetition tasks, one with English consonant letters and one with Hebrew consonant letters. We used the common procedure of verbal Hebb repetition learning (involving a single repeated Hebb sequence) adopted in most recent research on Hebb learning, which is based on Hebb's (1961) original work. Participants were retested on the same tasks after a period of about six months. This design provided multiple tests of the reliability of the Hebb learning measure. First, we examined whether Hebb repetition learning performance using Hebrew letters correlates with performance using English letters within a given session (i.e., parallel tests reliability). Second, we tested whether performance in the Hebb repetition task at initial testing (T1), predicts performance at the retest (T2) (i.e., test-retest reliability). To preview our findings, the reliability of individual Hebb repetition learning performance in Experiment 1 across stimuli and in time was close to zero. We further aimed to

replicate these disturbing findings in *Experiment 2* with a different population (French speakers), and with a different Hebb task. In this experiment we measured the concurrent learning of two repeated sequences, aiming to improve reliability scores. Similar to Experiment 1, Experiment 2 produced very low scores of test-retest reliability. Admittedly, this initial investigation did not systematically examine the many possible variations of the verbal Hebb tasks as outlined above. Nevertheless, at least with the task variant used here, participants' performance was found to be unreliable.

EXPERIMENT 1

Method

Participants

Forty-seven students at the Hebrew University (13 males, M age 24.68, SD = 2.36) participated in the first session of the study. Thirty of them successfully completed both test sessions (9 males, M age 24.40, SD = 2.66). Participants were all native Hebrew speakers with a high proficiency in English (highest proficiency score on the English University exam) and they were paid for participation.

Materials and Procedure

Hebb repetition task

The procedure of the Hebb repetition task was based on the one described by Page et al. (2006)[1]. Sequences of eight consonants were presented visually for immediate serial

[1] We selected for our study the task used by Page et al. (2006), because it has a simple and straightforward procedure, because it was shown to produce a strong HRE at the group-level, and because of its central role in

recall. One particular sequence, the Hebb sequence, was repeated every third trial. The unrepeated sequences, i.e., filler sequences, acted as the control condition against which Hebb sequence performance was measured. The materials comprised two blocks, each containing 3 practice and 36 experimental trials (12 repetitions of the Hebb sequence, 24 filler sequences). Each participant performed one block with English letters and one block with Hebrew letters, with the participants allocated equally and randomly to each of the two block orders. The letters used in the English block were the following consonant letters (Z, R, T, P, S, D, F, G, H, J, K, L, M, C, V, B, N). The letters in the Hebrew block were the following consonant letters [ר, ש, ת ב, ג, ד, ז, ח, ט, כ, ל, מ, נ, ס, ע, פ, צ, ק]. No letter was repeated in a given trial, and no sequence was repeated other than as part of the Hebb repetition manipulation. Importantly, three-letter alphabetic runs (e.g., B, C, D/ג, ב, א) were not permitted nor were consonant sequences that formed legal Hebrew consonantal roots[2]. With the constraints given above, 10 unique sets of sequences (each set containing one repeated Hebb sequence and 24 unrepeated filler sequences) were constructed for English and for Hebrew, and participants were randomly assigned to two sets at T1 (one with English consonants, one with Hebrew consonants), and to two different sets at T2. There was a time gap of about 6 months (M: 179.3 days, SD: 13 days) between T1 and T2. Figure 1 shows an example of a number of possible trials.

On each trial, the eight consonants were presented for 500 ms with an inter-stimulus interval of 0 ms. Immediately after presentation, a recall screen showed the eight consonants,

---

demonstrating Hebb repetition learning. Additionally, the use of consonant sequences in this task allows for simple adaptation across language conditions.

[2] In Hebrew consonantal roots are used to form words by adding vowels or transfixes to the root itself. Usually these roots consist three to four constants that are also words in Hebrew (see Frost, Deutsch & Forster, 1997).

[3] Note that we obtained qualitatively identical results when using "edit scoring", a scoring method based on calculating the smallest number of operations (insertion, deletion, or substitution of a single character) needed to modify the recalled sequence so it matches the presented sequence (i.e. Levenshtein distance, Levenshtein, 1966), then subtracting this number from the list length (8).

arranged in a circle around a central question mark. Participants were instructed to recall the order of the consonants by clicking the items in the order of presentation and to click the question mark for omitted consonants. Note that the positioning of the letters around the question mark was random on each trial, preventing a visuospatial recall and/or learning strategy. After the participant had clicked eight responses, he/she was able to advance to the next trial by pressing the spacebar. This clicking response format has the advantage (over immediate *verbal* serial recall) of allowing for automatic response registration and avoids the disadvantage of intrusion of items that were not presented (see also Bogaerts et al., 2016; Page et al., 2006; Szmalec et al., 2011).

(Figure 1 about here)

Measuring Hebb learning performance

In the Hebb task, an item is typically scored as correct if it was recalled in the correct position in the sequence.[3] Two main measures have been used in the literature to capture the improvement on the repeated Hebb sequence relative to performance on fillers (the HRE): the *Gradient measure*, and the *Halves measure* (see Bogaerts, Szmalec, Duyck, & Page, under review, for an elaborate discussion).

(1) *Gradient measure*: this common technique takes the gradient of the regression line through points representing the performance on successive Hebb repetitions and compares it with the gradient for corresponding filler trials, for each individual participant (e.g., Page et al., 2006; Gould & Glencross, 1990; Mosse & Jarrold, 2008; Archibald & Joanisse, 2012;

[3] Note that we obtained qualitatively identical results when using "edit scoring", a scoring method based on calculating the smallest number of operations (insertion, deletion, or substitution of a single character) needed to modify the recalled sequence so it matches the presented sequence (i.e. Levenshtein distance, Levenshtein, 1966), then subtracting this number from the list length (8).

Szmalec et al., 2011).

(2) Halves measure: this measure, put forward in several developmental Hebb learning studies (e.g., Mosse & Jarrold, 2008; Archibald & Joanisse, 2012; Smalle et al., 2015) captures the divergence in performance across Hebb repetition trials compared with filler trials, by collapsing the trials of each sequence type into first and second half scores and comparing the learning in terms of improvements across the two halves of the task, first vs. second.

In addition to these traditional measures, a recent alternative measure of the Hebb effect is derived from mixed logit models (see Bogaerts et al., 2016, for an application of this analysis method to the Hebb repetition paradigm). The degree of Hebb learning for a given subject is measured by the individual's coefficient of the interaction depicting the different effects of repetition for Hebb vs. filler trials. As we do not have a particular stance regarding the preferred measure of Hebb repetition learning, in the present study we assessed the reliability of all three measures. We contrasted the reliability of these Hebb learning indices with the reliability of overall serial recall performance of individual participants as measured by their average scores on unrepeated filler trials.

Results

Figure 2 shows the learning curves for English and Hebrew consonant sequences, at T1 (initial testing) and T2 (retest). The figures show a clear learning effect of repeated Hebb sequences across languages and sessions, relative to unrepeated filler sequences. This concurs with all previous studies reporting a HRE. To assess the statistical significance of Hebb learning on the group-level, we conducted analyses for both Hebrew and English, for

the initial test as well as the subsequent retest, across all three Hebb learning measures. First, the gradient values were entered into an analysis of variance (ANOVA) with Sequence type (filler vs. Hebb) as the independent variable, and we observed a significant main effect of Sequence type for all four tasks (Hebrew/English, test/retest), indicating a systematic significantly larger learning slope for the repeated Hebb sequences. For the second Hebb learning measure, the halves scores were entered into an ANOVA with Sequence type (filler vs. Hebb) and Halves (first six presentations vs. last six) as independent variables. Again we observed a significant interaction between Sequence type and Halves, demonstrating a divergence in performance across the two parts of the experiment on the repeated Hebb sequence compared to unrepeated filler sequences. Finally, we ran a logistic mixed effect model with accuracy as the dependent variable, fixed effects for Sequence type, Presentation (1-12) and their interaction, as well as by-subject and by-item intercepts and by-subject slopes for Sequence type and the Sequence type by Presentation interaction[4]. Again, a significant fixed effect of the interaction (Sequence type by Presentation) confirms the presence of a HRE in the sample. The summary of these analyses is presented in Table 1. Together, these analyses show that all measures result in a robust Hebb effect at the group-level, across sessions and across languages.

(Figure 2 about here)

(Table 1 about here)

[4] The models employed included the fullest random effects structure justified by the design that still allowed the model to converge. As a modeling procedure, when the full random model did not converge, we removed random terms that were not of theoretical interest, in this case for example the main effect of Presentation (Barr, Levy, Scheepers, & Tily, 2013).

The Hebb task: reliability characteristics

We now turn to the primary aim of the study, assessing the reliability of individual performance scores. In the upper panel of Table 2 we report the split-half reliability of the different task measures. However, as mentioned previously, this type of reliability taps only the internal consistency of the task's measures and not the stability of measures in time and across testing materials. Our design enabled us to assess, on the one hand, the reliability of overall individual capacity in serial recall, and, critically, on the other hand, it provided us with two independent measures of the reliability of participants' ability *to learn through repetition* (the HRE). First, participants' performance was compared within sessions between materials (English/Hebrew, i.e., parallel tests reliability), and second, it was compared across sessions within materials (i.e., test-retest reliability), once for Hebrew, and once for English.

Overall serial recall capacity: Here we asked whether performance with Hebrew filler sequences is correlated with performance with English filler sequences, and whether performance on fillers in the initial test (T1) is correlated with performance in retest (T2). Performance on fillers is an index of short-term memory span and does not reflect the ability to learn from repetition.

Individual Hebb learning ability: Here we asked (1) whether individual Hebb repetition learning performance with Hebrew letters predicted individual learning performance with English letters, and (2) whether individual learning performance in Hebrew and English at T1, predicted individual learning performance in Hebrew/English at T2. The results are presented in Table 2, and respective scatterplots in Figure 3.

(Figure 3 about here)

As can be seen in Table 2, the results of our three operational measures of the tasks' reliability (split-half, parallel tests and test-retest) show a very similar pattern: Individual overall serial recall capacity, as measured by mean filler performance, has a high reliability coefficient of around .80 (for comparison, reliability scores of standard cognitive tests are typically about .70 or more). In sharp contrast, for the same sample of participants, the three measures of Hebb learning, the gradients of improvement, the halves, and the coefficient measure, showed a very low level of reliability. There is some correlation between odd versus even trials within the same language condition and session (specifically for the Halves measure that reaches a corrected split-half coefficients of .60), suggesting that tests within language display some internal consistency. However, the correlations of individual performance across time or materials are near-zero.

(Table 2 about here)

Discussion

The results of Experiment 1 clearly show that whereas overall serial recall performance as measured in the Hebb repetition task is a stable and reliable capacity of an individual, the *learning* from repetition is not. First, in line with previous reports (Mosse & Jarrold, 2008), we observed moderate levels of (within-language, within-session) split-half reliability. By contrast, bilinguals' Hebb learning performance with letters in one language did not predict their Hebb learning performance with letters in another language. Moreover, within any language, their performance in one testing session did not correlate with their performance in a subsequent session. Note that these findings are independent of how the

HRE is measured; by the gradient measure, the halves measure, or the individual's coefficient extracted from a logit mixed model.

It should be noted that Mosse and Jarrold (2008) opted to look at the predictive value of individual's Hebb repetition performance by correlating Hebb performance on the second half of the task partialling out performance on the first half, thereby avoiding gradient and difference scores, which have been argued to be inherently less reliable (see Carter, Krause, & Harbeson, 1986; Dunlap, Kennedy, Harbeson, & Fowlkes, 1989; Mosse & Jarrold, 2008, and an extended discussion of this issue in the General Discussion). We followed this procedure as well, and estimated the reliability of such partial *r* measure. However, again, both within- (r = .03) and between-session correlations (English: r = -.15 Hebrew: r = -.07) remained close to zero and nonsignificant.

Taken together, the results of Experiment 1 raise serious doubts whether Hebb repetition learning performance as revealed in the Hebb task, reflects a stable ability of an individual, and can thus serve as a reliable predictor of other cognitive capacities. However, before reaching this conclusion, we conducted a second experiment. The aim of Experiment 2 was to investigate whether an alternative Hebb learning task (with an increased number of repeated Hebb trials) displays improved psychometric properties.

EXPERIMENT 2

In the typical Hebb paradigm (and the one we have used in Experiment 1) there is but one single repeated Hebb sequence, with 12 Hebb trials across the experimental session, and learning is assessed given participants' performance in the final trials relatively to the initial trials. Psychometric considerations in individual differences studies suggest, however, that a

larger number of trials would reduce measurement error and increase the task's sensitivity (see also Siegelman, Bogaerts, & Frost, 2017). In addition, since all measures of learning in the Hebb paradigm are based on the increase in performance from the few first Hebb trials to the few last Hebb trial, spurious high performance in the initial trials would inevitably result in a low learning score. A possible experimental approach to alleviate these problems, at least to some extent, is to have each individual learn more than one Hebb sequence. Indeed, recent studies demonstrated that participants are able to learn two different Hebb sequences concurrently (see Saint-Aubin, Guerard, Fiset, & Losier, 2015; Hitch, Flude, & Burgess, 2009). In Experiment 2 we employed such procedure. This enabled us to achieve three important goals. First, to launch a constructive replication of Experiment 1, using a different experimental design, and testing a different population of participants. Since the task employs letters, which are linguistic stimuli, testing the task in yet another language strengthens our findings. Second, to provide us with yet another measure of the task reliability, by correlating performance of participants in two lists within each testing session. Third, to examine whether a compound learning measure of two lists rather than a single one would result in increased test-retest reliability.

Method

Participants

Forty-six students at the University of Aix-Marseille (17 males, M age 21.30, SD = 4.05) participated in the first session of the study. Forty-four of them successfully completed both test sessions (15 males, M age 21.00, SD = 3.60). Participants were paid for

participation. Aside from two highly proficient bilinguals, all participants were native French speakers.

Materials and Procedure

Dual-list Hebb repetition task. The procedure of the verbal Hebb task was similar to that of Experiment 1 except that now *two* particular sequences were repeated every four trials. Repeated sequences were always preceded by an unrepeated filler sequence. Thus, if the first repeated sequence is referred to as HebbA and the second as HebbB, the task was constructed as follows: filler=>HebbA=>filler=>HebbB=>filler=>HebbA=>filler=>HebbB (see Saint-Aubin et al., 2015, for a similar procedure). The task contained then 48 experimental trials (12 repetitions of Hebb sequence A, 12 repetitions of Hebb sequence B, 24 filler sequences), plus 3 practice trials. As in the "English" condition of Experiment 1 the letters were taken from the full set of consonants, with the exception of W, Y and Q. Using the same constraints as those outlined previously, five different sets of sequences were constructed (each set containing two non-overlapping Hebb sequences and 24 unrepeated filler sequences). Every set had two versions, in which the order of Hebb sequence A and B were swapped, thus creating 10 different sets in total. Within every set, the filler sequences consisted of four items from each of the Hebb sequences. Participants were randomly assigned to a set at T1 (initial testing) and a different set at T2 (retest). The two testing sessions were about one moth apart (M = 26.78 days, SD = 4.71)[5].

Results

[5] We opted for this shorter between-session interval to verify that the low reliability estimates in Experiment 1 are not due to the relatively long six-month interval employed.

Figure 4 shows the learning curves at T1 and T2. The figure shows that participants managed to learn the two Hebb sequences simultaneously, both at T1 and at T2. To assess the statistical significance of the HRE on the group-level, we conducted analyses for the initial test as well as the subsequent retest, again, across all three Hebb learning measures (i.e., the gradient measure, halves measure, and the logistic mixed effect analysis, see Table 3).

First, we computed, for each participant, the gradient of improvement across the 12 repetitions for repeated sequence A, B, and their associated (preceding) nonrepeated filler trials. An ANOVA with Sequence type (filler vs. Hebb) and List (A vs. B) revealed that the gradient of improvement was significantly higher for the repeated sequences than for the nonrepeated sequences. The interaction between Sequence type and List was not significant ($F < 1$), indicating a comparable learning rate for the two repeated Hebb sequences. Second, halves scores were entered into an ANOVA with Sequence type (filler vs. Hebb), Halves (first six vs. last six presentations) and Lists (A and B). A significant interaction between Sequence type and Halves, in the absence of an interaction with List ($F \leq 1$), demonstrates a divergence in performance across the two parts of the experiment for both repeated Hebb sequences compared with their associated unrepeated filler sequences. Finally, we ran a logistic mixed effect model with accuracy as the dependent variable, fixed effects for Sequence type, Presentation (1-12), Lists (A and B) and the Sequence type:Presentation interaction as well as the three-way Sequence type:Presentation:List interaction, as well as by-subject and by-item intercepts and by-subject slopes for Sequence type, Presentation and their interaction. A significant fixed effect of the interaction (Sequence type by Presentation) confirms yet again the presence of a HRE in the sample.

(Figure 4 about here)

(Table 3 about here)

The dual-list Hebb task: reliability characteristics

Whereas it is of course crucial to show that Hebb learning took place in our task at the group-level, the primary aim of Experiment 2 was, again, to assess the *reliability* of individual performance scores in the *dual-list Hebb task*. In line with the analysis of Experiment 1 we calculated reliability estimates of individual performance in general serial recall, and then of the HRE. Reliability of the HRE was assessed first within session between Lists (A/B), and then across sessions collapsed over A- and B-lists. We expected that the use of two repeated sequences would potentially lead to improved reliability.

As a first step, we tested whether performance on *filler* sequences associated with List A is correlated with performance with on *filler* sequences associated with List B (note that this corresponds to split-half internal consistency of *serial recall capacity*). Second, we tested whether performance on fillers in the initial test (T1) is correlated with filler performance in retest (T2) (this corresponds to the test-retest reliability of overall serial recall capacity of an individual). Critically, to evaluate the reliability of *Hebb learning* as an individual ability, we asked whether Hebb repetition learning performance on List A predicted learning performance on List B, and whether overall learning performance at T1 (across both lists), predicted overall individual learning performance at T2. These reliability coefficients are presented in Table 4, and respective scatterplots in Figure 5.

(Table 4 about here)

(Figure 5 about here)

The results of our two tests of the tasks' reliability (see Table 4) replicate the results of Experiment 1. Whereas individual serial recall capacity (measured by mean filler performance) displayed high reliability coefficient of around .70, between-list reliability and the test-retest based on each of the lists for the three measures of Hebb learning (the gradients of improvement, the halves, and the coefficient measure) showed again strikingly low reliability. The between-list finding is important, because the low test-retest reliability between sessions could, in principle, be attributed to the impact of task repetition (performance in the second session introduces additional variance related to experience with the task). If this were the case, however, a high correlation between two Hebb lists *within* a session would have been observed. This is not what the data, summarized in Table 4 (panel A), suggest. Notably, the lower row of Table 4 (panel B) displays the test-retest reliability estimates for the Hebb learning measures calculated across lists A and B. The measures are therefore based on a larger number of Hebb trials, which –so we hypothesized- would reduce the measurement error. The values after outlier removal are indeed somewhat higher (ranging from .21 to .36) but fall very far of acceptable reliability standards of psychological measurements[6].

Experiment 2 thus demonstrates that increasing the number of observations of Hebb trials in a dual-list Hebb task only minimally improves its test-retest reliability. Similar to Experiment 1, overall serial recall performance as measured in the Hebb task was found to be a stable and reliable capacity of an individual, whereas the ability to learn from repeated sequences in the Hebb repetition paradigm, the HRE, is not.

[6] See, for example, *Standards for Educational and Psychological Testing*, 2014.

General discussion

Hebb repetition learning has been the focus of a series of recent studies that consider the ability to learn sequences from repetition as an important theoretical construct. Verbal Hebb repetition learning thus specifically targets the assimilation of repeated phoneme/letter sequences that form spoken or printed words. The claim that the ability to learn sequences underlies a range of language (dis)abilities is rooted in a theoretical framework that considers most linguistic material to be recurrent sequences of small building blocks (such as phonemes, syllables, or letters, Page & Norris, 2009; Szmalec et al., 2009, 2012). The present study and our obtained results do not challenge this theoretical framework. Indeed, at the group-level, while contrasting performance of normal controls to that of individuals with a reading disability or SLI patients, the learning performance of normal controls in the Hebb repetition task has been shown to exceed that of clinical populations (Bogaerts et al., 2015; Gould & Glencross, 1990, Szmalec et al., 2011; Hsu & Bishop, 2014).

Our present results, however, call for caution when investigating individual differences in Hebb learning performance. That is, whereas the Hebb repetition task seems to produce systematic learning effects at the group-level, the extent of learning in this task seems to be a very poor proxy of *an individual's* learning ability. Indeed, we observed a clear group-level HRE in all versions of the task (English, French, and Hebrew materials, classical procedure and dual-list) and across different samples. In the same vein, overall serial recall scores as revealed by the task, exhibited strong reliability, in both parallel tests and test-retest. In contrast, individual HRE exhibited close to zero reliability, whether assessed by using more than one Hebb list within a single session, or through test-retest between sessions (and

importantly, it did not matter how the HRE was measured).

The present findings could then lead to one of two possible theoretical conclusions. First, that perhaps learning from repetition is not a stable and reliable individual ability. Second, that assessing this ability in the Hebb repetition task, by contrasting repeated with unrepeated sequences, results in a learning measure that is unreliable. Although we cannot dismiss the first possibility, the relatively large amount of group studies showing that performance in the Hebb repetition task is related to language disabilities, suggests that serial-order learning capacity has substantial theoretical validity. It seems then more likely that the low reliability estimates of the Hebb repetition task are related to its inherent poor psychometric properties. The important theoretical contribution of the HRE is to isolate the ability to learn from repetition from overall short-term memory capacity. However, to do so one has to revert to difference scores of slopes or mean performance. Difference scores measuring cognitive skills, although widely applied, typically suffer from low reliability (e.g., see Carter et al., 1986; Dunlap et al., 1989, for extensive discussions). This is partly due to the substantial shared variance between performance in the baseline and the experimental conditions (in our case, between the filler- and Hebb trials), which is extracted while computing the difference score (Rodebaugh et al., 2016).

In addition, the measure of learning in the Hebb repetition task is exceedingly fragile, because it is based on a too low number of observations, and it can be easily masked by a (spurious) high performance in the first trials [7]. Indeed, Experiment 2 showed that

[7] If an individual accidently scores high in one of the first trials then there will be relatively little room for him/her to improve across repetitions of the Hebb sequence and this will result in a spuriously low learning score (see also Staels & Van den Broeck, 2014). Note that the same negative relationship between initial sequence recall performance and learning scores holds for "true" high initial scores (i.e., individuals with a high serial short-term memory capacity), however, as serial recall performance is stable over time and materials this is a problem of task *validity* rather than *reliability*.

increasing the number of observations by introducing two Hebb lists to learn instead of one, somewhat improves the task's reliability, albeit not to the level of a tool that has predictive validity. Since reliability is a necessary condition for predictive validity, the joint findings of our two experiments clearly demonstrate that Hebb repetition performance, as it is commonly measured, has limited potential as an individual measure and is unlikely to make reliable predictions of individual differences in linguistic abilities.

An emerging question given the present findings is how to account for previous studies that did report correlations between linguistic performance and Hebb repetition learning (Mosse & Jarrold, 2008; Bogaerts et al., 2015). Whereas it is obviously possible that these findings originate from Type I error (and are, therefore, spurious correlations), the replication of findings across research groups, which are in line with most group findings, make this possibility perhaps less likely. Note that the problem with low-reliability of the Hebb repetition task is in fact a double-edged sword. Since the correlation between two measures is upper-bounded by their reliability ($\rho_{xy} \leq \sqrt{\rho_{xx} * \rho_{yy}}$), a weak correlation between a poorly reliable Hebb learning measure and a presumably more reliable linguistic measure could in fact reflect a stronger true correlation. Thus, only a psychometrically reliable task would accurately reveal the theoretical link between Hebb repetition learning and linguistic skills.

Methodological considerations and future directions

As outlined in the introduction, the many Hebb repetition learning studies share the typical procedure in which a single Hebb list is presented for immediate serial recall on eight to twelve occasions, each separated by non-repeated filler lists. The task variants used in the

literature vary however widely in their specific parameters (e.g., stimulus material, presentation modality, response format, etc.). Naturally, all of these parameters could potentially influence performance (e.g., Szmalec et al., 2011; Zimgibl & Koch, 2002) and potentially also task reliability. In the current investigation we have evaluated the psychometric characteristics of a visual Hebb task, employing a clicking response format that has been used in multiple recent Hebb studies in adults (e.g., Page & Norris, 2006; Szmalec et al., 2011), and in recent work focusing on individual differences (Bogaerts et al., 2016). Our findings do not preclude the possibility that Hebb repetition tasks with other parameters (e.g., auditory presentation, verbal immediate serial recall as a response procedure, etc.) could perhaps fare better in terms of their psychometric properties. This, however, would require additional investigation.

Regarding the development of more reliable measures for tracking individual capacities, even further increasing the number of data points (e.g., using several Hebb lists, see Hitch et al., 2009) and/or adapting the task to each individual's memory span by changing the number of items in the Hebb sequence (see e.g., Hsu & Bishop, 2014) could be fruitful directions to investigate. It’s worth noting that although the low test-retest reliability of individual learning performance does not undermine the theoretical validity of the task in assessing serial-order learning on the group-level (e.g., Szmalec et al., 2012; Hsu & Bishop, 2014), even group-level studies, would benefit from more reliable proxy of the HRE. More reliable measures lead to decrease measurement error and thus provide increased power in detecting true effects.

In sum, Hebb repetition learning offers illuminating perspectives for understanding memory for serial order, language learning, and their interactions. This imposes a challenge

how to tap this theoretical construct so that it can be measured reliably, withstanding tests of psychometric validity.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Archibald, L.M.D., & Joanisse, M.F. (2012). Atypical neural responses to phonological detail in children with developmental language impairments. *Developmental Cognitive Neuroscience*, *2*(1), 139–151.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bogaerts, L., Szmalec, A., Hachmann, M. W., Page, M. P. A., & Duyck, W. (2015) Linking memory and language: Evidence for a serial-order learning impairment in dyslexia. *Research in Developmental Disabilities*, *43–44*, 106–122

Bogaerts, L., Szmalec, A., De Maeyer, M., Page, M. P. A., & Duyck, W. (2016). The involvement of long-term serial-order memory in reading development: A longitudinal study. *Journal of Experimental Child Psychology, 145, 139-156.*

Carter, R. C., Krause, M. & Harbeson, M. M. (1986). Beware the reliability of slope scores for individuals. *Human Factors, 28(6)*, 673-683.

Cumming, N., Page, M. P. A., & Norris, D. (2003). Testing a positional model of the Hebb effect. *Memory, 11(1)*, 43–63.

Couture, M., & Tremblay, S. (2006). Exploring the characteristics of the Hebb repetition effect for visuo-spatial information. *Memory and Cognition, 34*, 1720-1729.

Dunlap, W. P., Kennedy, R. S., Harbeson, M. M. & Fowlkes, J. E. (1989). Problems of individual different measures based on some componential cognitive paradigms. *Applied Psychological Measurement, 13(1)*, 9-17.

Frost, R., Forster, K.I., & Deutsch, A. (1997). What can we learn from the morphology of Hebrew: a masked priming investigation of morphological representation. *Journal of Experimental Psychology: Learning Memory, & Cognition, 23*, 829-856.

Gannon, S., Forster, K. F., Turcotte, T., & Jongenelis, J. (2001). Involvement of the hippocampus in implicit learning of supra-span sequences: The case of SJ. *Cognitive neuropsychology, 21(8)*, 867–882.

Gould, J.H., & Glencross, D.J. (1990). Do children with a specific reading disability have a general serial-order deficit? *Neuropsychologia, 28*, 271-278.

Hebb, D. (1961). *Distinctive features of learning in the higher animal*. In J. F. Delafresnaye (Ed.), *Brain mechanisms and learning* (pp. 37–46). Oxford, UK: Blackwell.

Hitch, G. J., Flude, B. & Burgess, N. (2009). Slave to the rhythm: Experimental tests of a model for verbal short-term memory and long-term sequence learning. *Journal of Memory and Language, 61*, 97-111.

Hsu, H. J., & Bishop, D.V.M. (2014). Sequence-specific procedural learning deficits in children with specific language impairment. *Developmental Science, 17*, 352–365.

Jefferies, E., Bott, S., Ehsan, S. & Lambon, R.M.A. (2011). Phonological learning in semantic dementia. *Neuropsychologia, 49* (5), 1208-1218.

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

Mosse, E.K., & Jarrold, C. (2008). Hebb learning, verbal short-term memory, and the acquisition of phonological forms in children. *The Quarterly Journal of Experimental Psychology, 61,* 505-514.

Mosse, E.K., & Jarrold, C. (2010). Searching for the Hebb effect in Down syndrome: evidence for a dissociation between verbal short-term memory and domain-general learning of serial order. *Journal of Intellectual Disability Research, 54(4)*, 295-307.

Page, M.P.A., Cumming, N., Norris, D., Hitch, G., & McNeil, A. (2006). Repetition learning in the immediate serial recall of visual and auditory materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 716–733

Page, M. P .A. & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word-forms. *Philosophical Transactions of the Royal Society B, 364,* 3737-3753.

Parmentier, F. B. R., Maybery, M. T., Huitson, M., & Jones, D. M. (2008). The perceptual determinants of repetition learn- ing in auditory space. *Journal of Memory and Language*, *58*, 978–997.

Rodebaugh T. L., Scullin R. B., Langer J. K., Dixon D. J., Huppert J. D., Bernstein A., Zvielli A., Lenze E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology, 125*(6), 840-851.

Saint-Aubin, J., Guérard, K., Fiset, S. et Losier, M.-C. (2015). Learning multiple lists at the same time in the Hebb repetition effect. *Canadian Journal of Experimental Psychology, 69*, 89-94.

Sechler, E. S., & Watkins, M. J. (1991). Learning to reproduce a list and memory for the learning. American Journal of Psychology, 104, 367–394.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81*, 105–120.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods, 49*, 418–432.

Smalle, E.H.M., Bogaerts, L., Simonis, M., Duyck, W., Page, M.P.A., Edwards M.G., Szmalec, A. (2015). Can chunk size differences explain developmental changes in lexical Learning? *Frontiers in Psychology, 6(1925)*,1-14.

Staels, E., & Van den Broeck, W. (2014). No Solid Empirical Evidence for the SOLID (Serial Order Learning Impairment) Hypothesis of Dyslexia. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000054

Szmalec, A., Duyck, W., Vandierendonck, A., Barberá Mata, A., & Page, M. P. A. (2009). The Hebb repetition effect as a laboratory analogue of novel word learning. *Quarterly Journal of Experimental Psychology, 62*, 435-443.

Szmalec, A., Loncke, M., Page, M. P. A., & Duyck, W. (2011). Order or disorder? Impaired Hebb learning in dyslexia. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1270-1279.

Szmalec, A., Page, M. P. A. & Duyck, W. (2012). The development of long-term lexical representations through Hebb repetition learning. *Journal of Memory and Language, 67,* 342-354.

Turcotte, J., Gagnon, S., Poirier, M. (2005). The effect of old age on the learning of supraspan sequences. *Psychol Aging, 20(2*), 251-60.

Zimgibl, C., & Koch, I. (2002). The impact of response mode on implicit and explicit sequence learning. *Experimental Psychology, 49(2)*,153-162.

Tables

Table 1. Experiment 1: Summary statistics and significance testing of learning in the group-level. Df (1, 46) for T1 and df (1, 29) for T2. For the Linear Mixed Model we report the Beta coefficient.

		Gradient			***Halves***			***Linear Mixed Model***	
		M (SE)	***F***	***p***	***M (SE)***	***F***	***p***	***β (SE)***	***p***
T1	***Hebrew***	.018 (.003)	28.65	*<.001*	.09 (.03)	13.09	*<.001*	.25 (.04)	*<.001*
	English	.015 (.004)	11.63	*=.001*	.09 (.03)	10.47	*<.01*	.13 (.03)	*<.001*
T2	***Hebrew***	.013 (.006)	4.38	*=.045*	.09 (.04)	5.34	*=.028*	.14 (.05)	*<.01*
	English	.025 (.006)	19.18	*<.001*	.15 (.03)	19.73	*<.001*	.24 (.05)	*<.001*

Table 2. Experiment 1: Reliability coefficients of the Hebb repetition task. (*p≤.05, ***p≤.001).

A. Split-half reliability: Correlations of performance for two halves (odd trials / even trials) of the same task at T1, for filler trials (*serial short-term memory capacity*) and for Hebb repetition measures (*Hebb learning ability*). Between brackets are Spearman-Brown corrected correlation coefficients.

	Overall serial recall (filler performance)	***Hebb Learning measures***		
		Gradient	*Halves*	*Coefficient*
Hebrew	.80*** (.89)	.20 (.33)	.33* (.50)	.29* (.45)
English	.81*** (.90)	.26* (.41)	.43** (.60)	.25* (.40)

B. Parallel testing reliability (within session between languages): Correlations of performance with Hebrew and English material at T1, for filler trials (*serial short-term memory capacity*) and for Hebb repetition measures (*Hebb learning ability*).

Overall serial recall (filler performance)	***Hebb Learning measures***		
	Gradient	*Halves*	*Coefficient*
.82***	.09	.22	.06

C. Test-retest reliability (between sessions): correlations of individual performance in the Hebb paradigm between the two testing sessions for Hebrew and English materials, for filler trials (*serial short-term memory capacity*) and for Hebb repetition measures (*Hebb learning ability*).

	Overall serial recall (filler performance)	***Hebb Learning measures***		
		Gradient	*Halves*	*Coefficient*
Hebrew	.78***	.28	.26	.12
English	.82***	-.26	-.20	-.03

Table 3. Experiment 2: Summary statistics and significance testing of learning in the group-level. df (1, 44) for T1 and df (1, 42) for T2. For the Linear Mixed Model we report the Beta coefficient.

	Gradient			***Halves***			***Linear Mixed Model***	
	M (SE)	***F***	***p***	***M (SE)***	***F***	***p***	***β (SE)***	***p***
T1	.015 (.004)	17.73	*<.001*	.08 (.02)	10.42	*<.01*	.05 (.01)	*<.001*
T2	.015 (.004)	15.19	*<.001*	.08 (.03)	8.52	*<.01*	.06 (.01)	*<.001*

Table 4. Experiment 2: Reliability coefficients of the verbal Hebb repetition task. (*p≤.05, ***p≤.001).

A. Within session between-list reliability (corresponds to split-half reliability): Correlations of performance on List1 and List2 at T1, for filler trials (*serial short-term memory capacity*) and for Hebb repetition measures (*Hebb learning ability*).

Overall serial recall (filler performance)	***Hebb Learning measures***		
	Gradient	*Halves*	*Coefficient*
.67***	.06	.20	.27*

B. Test-retest reliability (between sessions): correlations of individual performance in the Hebb paradigm between the two testing sessions for List1 and List2 as well as for performance collapsed across lists, for filler trials (*serial short-term memory capacity*) and for Hebb repetition measures (*Hebb learning ability*). Between brackets are values after the removal of outliers (see Figure 5).

Overall serial recall (filler performance)		***Hebb Learning measures***		
		Gradient	*Halves*	*Coefficient*
.69***	**List 1**	.02	.01 (.15)	.15 (.26*)
	List 2	.12	-.10 (.14)	.00 (.21)
	Collapsed	.05 (.29*)	.07 (.21)	.01 (.36*)

Figures

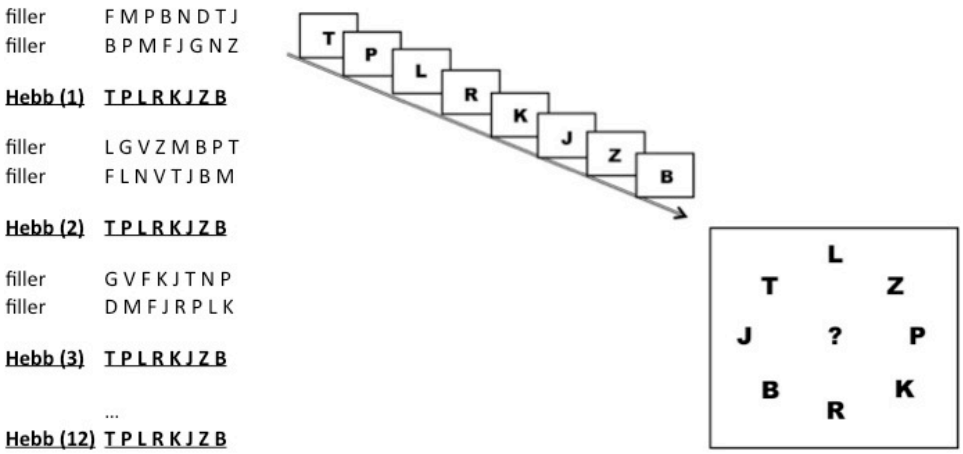


Figure 1. Depiction of the Hebb repetition task with English consonants.

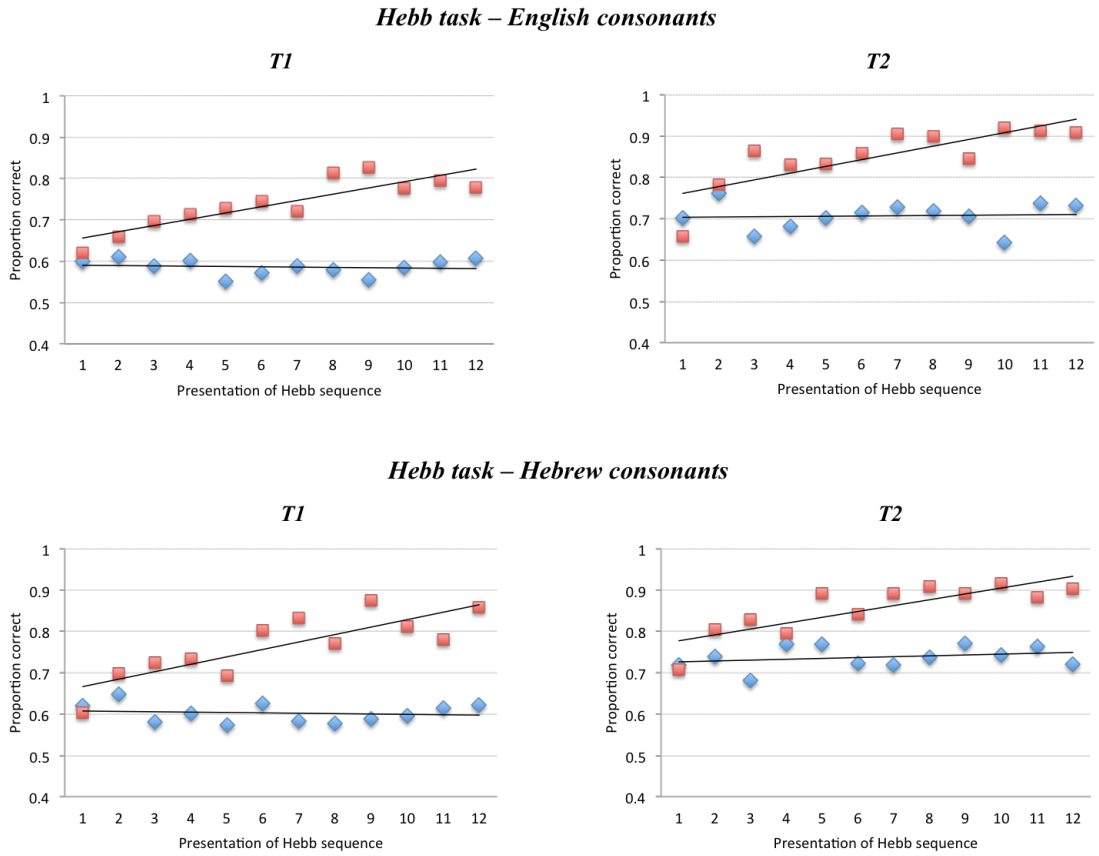
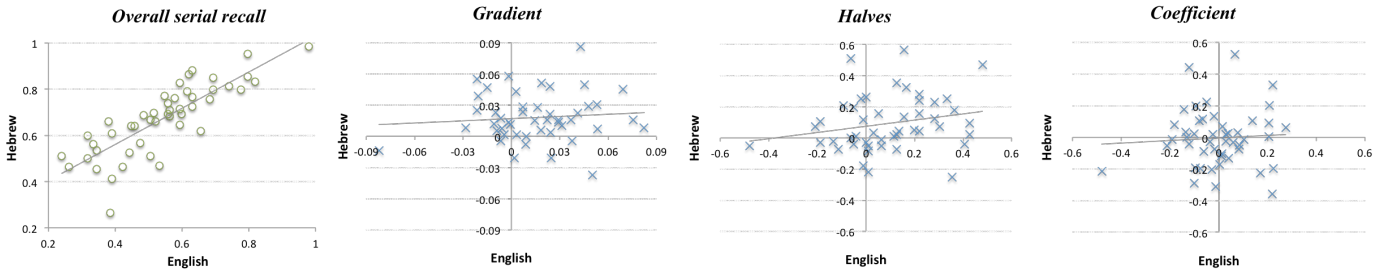


Figure 2. Experiment 1: Plots of the average proportion of correctly recalled items for Hebb (red) and filler (blue) as a function of presentation position of the Hebb sequence. Regression lines have been added to show the change in performance for repeated Hebb trials vs. filler trials.

A. Parallel testing reliability (within session between languages)



B. Test-retest reliability (between sessions)

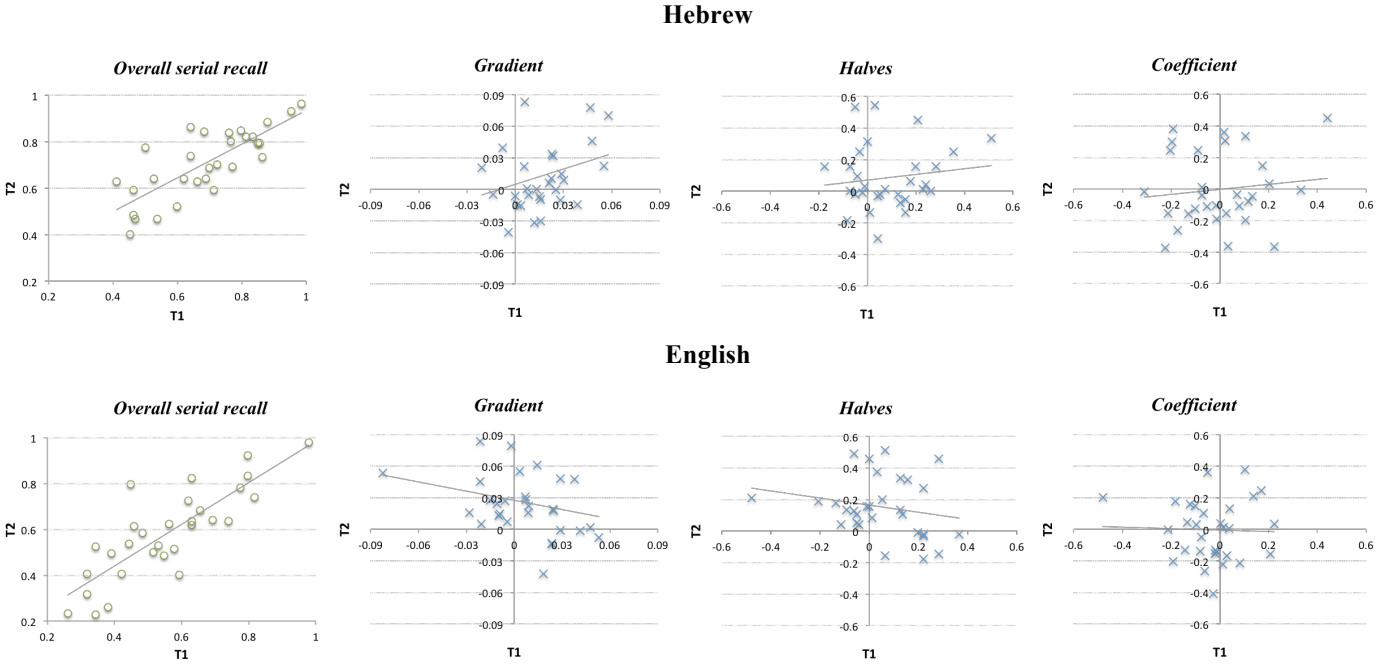


Figure 3. Experiment 1: Scatterplots for overall serial recall (green) on the one hand and the three learning measures (blue) on the other hand.

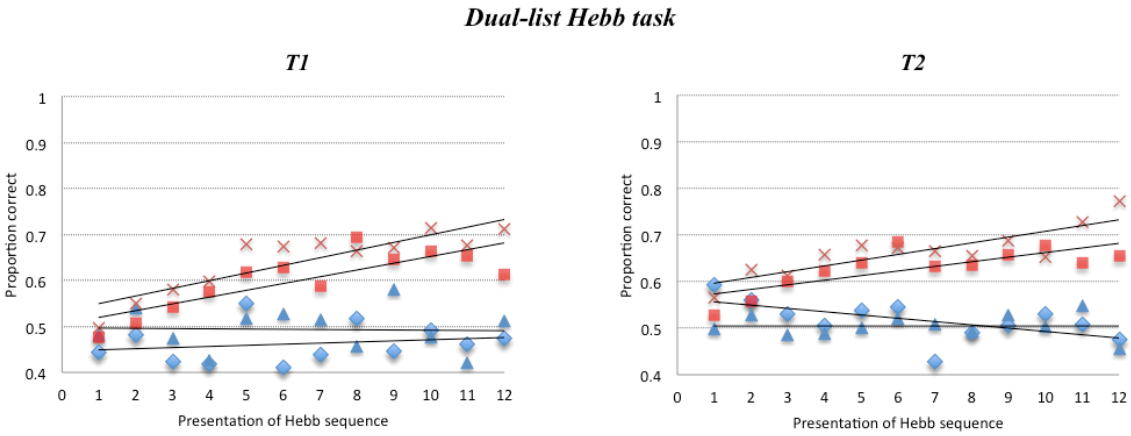
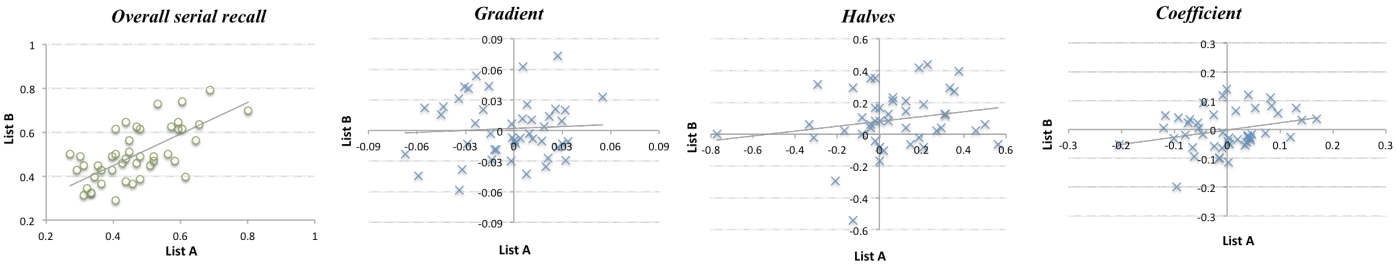


Figure 4. Experiment 2: Plots of the average proportion of correctly recalled items for Hebb (red), and filler (blue), as a function of presentation position. Squares and diamonds represent performance in List A, X-es and triangles represent performance in list B.

***A*. Within session between-list reliability (corresponds to split-half reliability)**



***B*. Test-retest reliability (between sessions)**

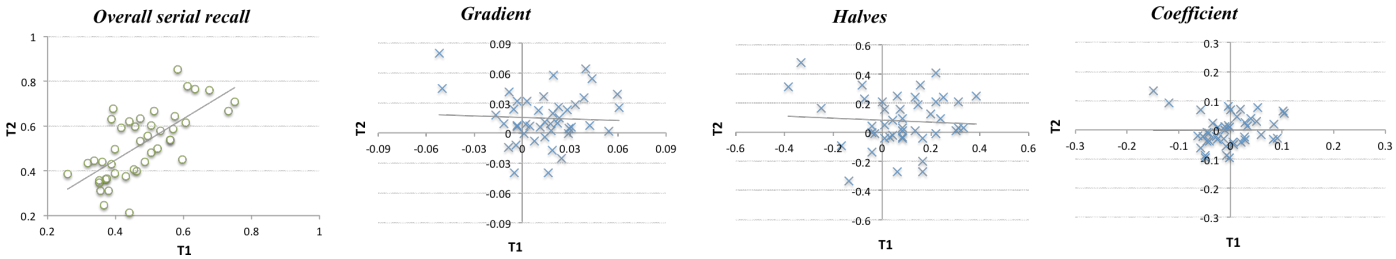


Figure 5. *Experiment 2:* Scatterplots for overall serial recall (green) on the one hand and the three learning measures (blue) on the other hand.