A novel application of generalizability theory to evaluate the reliability of the recognition memory test

Angela Medvedeva^{1,2} · Giulia Galli¹ · Aleksandra Petrovskaya³ · Oleg N. Medvedev⁴

Accepted: 7 February 2022
© The Author(s) 2022

Keywords Recognition test · Generalizability Theory · Reliability · Measurement · Memory · Assessment

Introduction

Episodic memory, a necessary function in everyday life, is defined as memory for an event or item and associated contextual details including time and place (Tulving, 2002). Although there are multiple tests to measure episodic memory in the laboratory, the predominant measures are recall and recognition tests (Otani & Schwartz, 2019; Renoult & Rugg, 2019). After a series of items is presented (pictures, words, or other stimuli), participants can be asked to list the items they remember (recall) or indicate which items were presented from a set of items that includes new items as well as the presented ones (recognition).

Reliability is defined as an estimate of the true measurement of a construct that accounts for both variability

and error across multiple independent measurements (Brennan, 2010; Cronbach, 1947). A reliable test is one that can estimate true scores between individuals over time and which has minimized error due to the measurement itself. For example, the interaction between the experimenter and the participant and different instructions for completing the test can potentially bias measurement, and scores can change between people because of greater measurement error in certain cases rather than actual differences in the construct. Common measures of reliability include test–retest reliability, which is merely a correlation between test scores at different time-points, and inter-rater reliability, a measure of the agreement between evaluators collecting observational data independently. It is important to ensure that a test is reliable and as accurate as possible across people, sessions, and time by identifying potential sources of error, including interactions between different factors, and minimizing them. In this way, we can estimate true differences between people and accurately measure effects of experimental manipulations, especially in memory tests.

Previous attempts to measure the reliability of episodic memory tests have been limited to neuropsychological tests. For example, Cheke and Clayton (2015) found no positive correlations between any of the episodic measures implemented, including free recall, and a non-linear relationship emerged only between two tests but not others. However, they did not include the recognition test among the measures of interest. Other studies have examined the reliability and validity of neuropsychological tests involving recognition, and these tests are typically used with patients. While these neuropsychological tests, including the California Verbal Learning Test, Auditory Verbal Learning, and Continuous Recognition test, have shown satisfactory reliability and validity (Larrabee & Curtiss, 1995), they are not frequently used in the laboratory with

Highlights

The psychometric properties of word recognition tests remain unknown.

Generalizability (G) theory is an underused, powerful reliability testing method.

We applied G theory to recognition test data with different trial numbers and languages.

The test showed high reliability in both languages with at least 40 trials.

✉ Oleg N. Medvedev
oleg.medvedev@waikato.ac.nz

¹ Department of Psychology, Kingston University, Penrhyn Road, Kingston Upon Thames KT1 2EE, UK

² Vivian L. Smith Department of Neurosurgery, University of Texas Medical School at Houston, Houston, TX 77030, USA

³ Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, 119991 Moscow, Russia

⁴ School of Psychology, University of Waikato, Hamilton 3240, New Zealand

experimental studies. The lack of use of neuropsychological tests is partly due to the large overlap with short-term memory (Hasselmo, 2011, p. 13–20) and partly due to inflexibility; stimuli within the tests cannot be easily changed, reduced, or expanded to meet the needs of the experiment without compromising validity or reliability. In contrast, experimental studies aiming to examine or manipulate a specific aspect of episodic memory change the content, length, and characteristics of the stimulus set while keeping the general recognition test structure. Moreover, studies can also implement varying delays between stimulus presentation and test, which is not always possible with neuropsychological tests that have fixed instructions and structure. Thus, a major question that remains to be answered is how the reliability of the “flexible” laboratory recognition test changes across different stimuli, number of test trials, populations, and languages. Will the test remain reliable and will measurements using the test be robust and comparable across studies?

Currently, the reliability of psychological and cognitive tests is predominantly examined using Classical Test Theory (CTT; Livingston, 1972; Miller, 1995), including Cronbach’s alpha that is merely an overall correlation between test items. CTT considers that any test score includes a true score (e.g. true differences between people) and an error, but it cannot differentiate between multiple sources of error such as individuals, items, language, and interactions between them that impact reliability. Generalizability theory (G theory) is an extension of CTT and goes beyond CTT’s limited assumption of considering error variance as a single factor (Allen & Yen, 1979) and can identify and evaluate specific sources of measurement error in a test (Cronbach et al., 1963). It is named *Generalizability Theory* because it estimates the extent to which the influence of any specific source of error variance can be generalized to all possible situations and contexts as opposed to only the specific testing situation for which the data had been obtained (Cronbach et al., 1963). G-Theory assesses numerous sources of variance contributing to the measurement error associated with the main variable of interest (Cardinet et al., 1976), such as a test score. In naturally occurring environments, there are more factors including personal (e.g., personality), methodological (e.g., psychometric characteristics of the measure used), and situational (e.g., time of the day) that might each independently contribute to measurement error. G-Theory provides an advanced method for assessing these factors and their interactions, thus contributing to the improvement of methodology and precision of an assessment (for review see Cardinet et al., 2010). G theory is superior to other methods of reliability assessment from CTT, such as test–retest and inter-rater reliability, because it is possible to estimate the unique impact of

multiple external (e.g. language) and internal (e.g. number of items) factors on measurement accuracy (Bloch & Norman, 2012). Thus, G theory can be applied to identify flaws in the recognition test, particularly those related to the flexible structure (different languages and number of items). G theory can examine whether the recognition test is biased due to the use of different words in different languages (e.g. each person receives a different set of words from a large pool in his or her native language – Russian or English).

It is important to establish the reliability of episodic measures to evaluate memory function and to study memory enhancement. Specifically, a growing number of studies in episodic memory have examined the possibility of enhancing memory function with non-invasive brain stimulation (NIBS), and many of these studies measured changes in memory performance using laboratory recognition tests. We pooled trial-by-trial data from the control group of multiple NIBS experiments in our lab (Medvedeva et al., 2019; Petrovskaya et al., 2020, in preparation) to accumulate a large sample size of Russian and English participants who completed the same memory test (identical stimuli) in Russian and English, respectively, for reliability testing.

The aim of the current study was to apply G Theory to establish the reliability of the recognition test and evaluate specific sources of measurement error associated with different languages and words used to assess memory performance. G theory was applied in two parts using a cross-sectional design with the individual as the object of measurement. Generalizability study (G-study) estimated sources of variance across individuals and trials in the current measurement design. Decision study (D-study) examined generalizability of recognition performance scores for different numbers of trials to establish reliability across trials and languages.

Method

Participants

The sample for this study was extracted from three experiments: (Medvedeva et al., 2019, Experiment 1) with an English sample (Medvedeva et al., 2019, Experiments 1–2, English Control 1) that was supplemented by a second experiment with an identical method (Medvedeva, 2019, Chapter 5 Experiment 4, English Control 2) and a third experiment with a Russian sample (Petrovskaya et al., 2020, Russian Control 1 and 2). The demographics for each sample are shown in Table 1; there were 80 individuals in total, 40 for each language (Russian and English) and there were no significant differences in age, $p = 0.07$.

Table 1 Demographics of samples included in the analysis

	Russian Control 1	Russian Control 2	English Control 1	English Control 2
Sex	12 F, 10 M	13 F, 5 M	14 F, 2 M	20 F, 4 M
Age (<i>M</i> , <i>SD</i>)	21.5, 4.08	19.12, 1.22	23.34, 5.00	20.96, 2.48
<i>n</i>	22	18	16	24
Original <i>n</i>	28	28	17	26

For sex, F denotes female and M denotes male. *n* reflects the sample size used in the current study, and Original *n* reflects the total usable (without outliers or issues) sample size from the completed experiment.

Procedure

The method and stimuli for all experiments followed the control group procedure detailed in Medvedeva et al., (2019) (Experiment 1) with the exception that for the Russian groups, stimuli were translated to Russian for the third experiment and participants from an additional Russian group were included (an active stimulation group that showed no difference in performance compared to the control group; see Petrovskaya et al., 2020). Together, the two English experiments and two Russian groups in the third experiment met sample size requirements ($n \geq 76$) for the reliability study with $(1-\beta)=0.85$ and $f \geq 0.25$ under $p=0.05$ (Atilgan, 2013). In each two-part experiment, participants judged the pleasantness (pleasant or not) of individually-presented words (MRC database; Coltheart, 1981) and memorized them on the first day (160 words) and 120 of these words were presented a day later in an old/new recognition task (indicate whether the item is old – previously presented – or new) along with 81 new words. See Appendix 1 for the pool of 240 words in Russian and English. For the current work, we only considered trial-by-trial performance on the 120 items that were presented at study and test. All experiments were approved by the authors' institutional ethics committees. The results of all experiments confirmed that the performance of the control group was not affected by the NIBS-specific procedure (application of electrodes, short duration of stimulation, etc.).

Data Analyses

Prior to the main analyses, normality of distribution was tested and ensured for both Russian (kurtosis = -0.28, skewness = -0.68) and English (kurtosis = -0.47, skewness = -0.59) groups. The two groups had significantly different average memory performance, $t(78) = -3.18$, $p = 0.002$, with higher performance for the English sample ($M = 0.75$, $SD = 0.13$) than the Russian sample ($M = 0.66$, $SD = 0.14$) but a medium effect size, $d = 0.66$. G theory analyses followed recommendations of Cardinet et al. (2010) and employed EduG 6.1-e software (Swiss Society for Research in Education Working Group, 2006). Accuracy data for each trial reflected correctly recognizing an

“old” word as previously-presented or incorrectly indicating that the word is “new” (1 = correct and 0 = incorrect or no response), and each participant had complete data for 120 trials. G theory was applied by defining the measurement design and estimating variance components in G-study, and conducting a D-study. The D-study involved calculating generalisability coefficients (G-coefficients) and variance estimates for different numbers of trials to examine the reliability of the test.

A person (P) nested within language (L) by trial (T) design expressed as P:L x T was implemented in which an equal number of participants ($n = 40$) was nested in each language (L) (Russian or English). Person nested in language (P:L) was the object of measurement and was not considered a source of error, while trial (T) was instrumentation facet because object of measurement (P) cannot be nested in the instrumentation facet (L). Therefore, a P:L/T $40 \times 2 \times 120$ measurement design was specified in EduG. Interactions between person and language may reflect an effect of language. A two-way ANOVA with a 2-level random effects measurement design defined as person nested in language by trial was conducted to calculate variance components due to person, language, trial, and interactions between them. EduG software accurately estimates variance components with application of Whimbey's correction coefficient (Cardinet et al., 2010) expressed as $(N(f) - 1)/N(f)$, with $N(f)$ representing population size of the (f) facet in the G-study design. G-study estimated variance components using standard ANOVA approach. D-study computed G coefficients using equations by Brennan (2010). G-coefficients reflect the overall reliability and generalizability of the test scores across persons and trials (Bloch & Norman, 2012; Medvedev et al., 2017). While the relative G-coefficient (Gr) only considers variance directly related to the object of measurement, the absolute G coefficient (Ga) accounts for other sources of variance such as trial that could affect absolute measurement indirectly, so it is a more conservative measure of reliability. D-study involved removing one of two languages as a level from the model to compare reliability and generalizability of the test in each language using G-coefficients. In addition, different number of trials including 20, 40, 60, 80 and 100 were examined to evaluate the reliability of the test.

Table 2 ANOVA estimates for individual nested within language by trial for the full sample (above) and English and Russian groups (below)

Source	SS	df	MS	Recognition variance components				
				Random	Mixed	Corrected ^a	%	SE ^b
L	20.720	1	20.720	0.004	0.004	0.002	0.9	0.004
P:L	162.129	78	2.079	0.016	0.016	0.016	7.5	0.003
T	25.700	119	0.216	0.000	0.000	0.000	0.1	0.000
L x T	24.830	119	0.209	0.000	0.000	0.000	0.2	0.001
P:L x T	1773.771	9282	0.191	0.191	0.191	0.191	91.2	0.003
English								
P	75.492	39	1.936	0.015	0.015	0.015	7.8	0.004
T	25.592	119	0.215	0.001	0.001	0.001	0.6	0.001
P x T	801.907	4641	0.173	0.173	0.173	0.173	91.6	0.004
Russian								
P	86.637	39	2.221	0.017	0.017	0.017	7.4	0.004
T	24.937	119	0.210	0.000	0.000	0.000	0.0	0.001
P x T	971.863	4641	0.209	0.209	0.209	0.209	92.6	0.004

^aCorrected components are calculated using Whimbey's correction to classical ANOVA estimates. ^bSE refers to the mixed effects.

Results

G study

Table 2 includes conventional ANOVA estimates computed for G-study including person (P) nested in language (L) by trial (T) and their interactions. The results of the G study are interpreted with a reference to the generalizability of a single trial. The large P:L x T variance component (0.191) reflects a large amount of variance (91.2% of the total variance). There is a modest amount of variability between persons within language groups (0.016 or 7.5%). The G study results for English and Russian presented in Table 2 are very similar, suggesting only negligible differences due to language. Therefore, it was unnecessary to conduct D study separately

for English and Russian sub-samples, and the full sample was used in D-study.

Table 3 includes D-study results and shows that after accounting for all sources of error, the recognition test has good generalizability across all English and Russian speakers with a high absolute G-coefficient (0.92). Variance components in D-study were computed for 120 where all error components from the G study have been divided by 120. Interaction between individual and trial nested in language reflects individual differences in remembering different words and explains the most error variance in memory performance. Language alone explains only 10% of the differentiation variance. Similarly, trial and the interaction between trial and language accounted for a negligible proportion of the absolute error variance (Table 3). Standard error values

Table 3 D-study estimates for the full sample

Source of σ^2	Differentiation σ^2	Relative σ^2	% relative	Absolute σ^2	% absolute
P:L	0.016	0.016	81.8	0.016	81.8
L	0.002	0.002	10.2	0.002	10.2
T				0.000	0.0
LxT		0.000	0.0	0.000	0.0
P:LxT		0.002	8.0	0.002	8.0
Total σ^2	0.018	0.002	100%	0.002	100%
		Relative SE: 0.040		Absolute SE: 0.040	
G relative	0.92				
G absolute	0.92				

Error components involving the trial (T) component have been divided by 120. Standard error (SE); standard deviation (SD); Variance components for the Person (P) nested in Language:(L)×Trial (T) design including interactions. Grand mean = 0.702; Standard error (SE) of the grand mean: 0.015.

Table 4 D-study estimates for the full sample and different number of trials

Facets	20 Trials		40 Trials		60 Trials		80 Trials		100 Trials	
	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%
P:L	0.016	58.1	0.016	70.3	0.016	75.5	0.016	78.5	0.016	80.4
L	0.002	7.3	0.002	8.8	0.002	9.4	0.002	9.8	0.002	10.0
T	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
LxT	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
P:LxT	0.010	34.7	0.005	21.0	0.003	15.0	0.002	11.7	0.002	9.6
SE	0.097		0.069		0.056		0.049		0.044	
Gr	0.65		0.79		0.85		0.88		0.90	
Ga	0.65		0.79		0.85		0.88		0.90	

Estimates include Coefficient G relative (Gr), Coefficient G absolute (Ga), variance components (in %), and for the Person (P)×Trial (T) including interaction. Numbers in bold signify acceptable reliability of the test ≥ 0.80 (Brennan, 2010; Cardinet et al., 2010).

were negligible, indicating the robustness of both relative and absolute G coefficients and in turn reflecting good reliability of the test.

D study variance components were produced by dividing the estimated variance components from the G study that include T (trial) by 20, 40, 60, 80, and 100, respectively (Table 4). We found insufficient reliability at 20 and 40 trials with relative and absolute G coefficient of 0.65 and 0.79, respectively (see Table 4). However, starting from 60 trials, reliability reached an acceptable level (Gr and Ga = 0.85). Furthermore, a non-linear increase in reliability was observed with an increasing number of trials: Gr and Ga of 0.88 for 60 trials and 0.90 for 100 trials, comparable to the overall reliability of the test with all 120 trials (Gr and Ga = 0.92). Moreover, the grand mean was not significantly different between trial designs (20–100 trials), ranging from 0.69 for 20 trials to 0.70 for 40, 60, 80, and 100 trials. For all trial groups except 20 trials, standard error values were relatively low, reflecting the robustness of both relative and absolute G coefficients and acceptable reliability of the test. We have also estimated G coefficients for each language to evaluate robustness if the number of trials is reduced to the critical cut-off point (e.g. 60, 40 and 20 trials). For the English version, both G-coefficients were at 0.61 and 0.76 with 20 and 40 trials respectively, and reached an acceptable level of 0.83 with 60 trials. Similarly, the Russian test version reached good reliability with 60 trials (both G = 0.82), while with 20 and 40 trials the reliability was below acceptable cut-off point of 0.80.

Discussion

The aim of the current study was to establish the reliability of a standard laboratory recognition test and identify possible errors due to language and number of trials

that could bias recognition measurement. The interaction between person and trial in language produced the largest amount of error, perhaps because different people interacted with different words uniquely, and after accounting for the error, the measurement error due to language was negligible. The results demonstrated that the reliability of the test and generalizability of test scores across samples populations depend on the number of trials used in the test irrespective of the language used. The highest reliability and generalizability of the test was achieved with 120 trials. Moreover, there was little measurement error due to different words (as represented by trial), so reliability in recognition measurement is not contingent upon using the same set of words. Thus, conducting the test in different languages is unlikely to significantly bias measurement of recognition, although the results can only be generalized to English and Russian younger adults.

The D study showed that by reducing the number of trials, reliability gradually decreased but was still at acceptable level with at least 60 trials overall and in each language. This indicates that the test users should consider including higher numbers of trials ($> = 60$) to ensure reliability of the test because with 80 trials the test was as reliable as one with 100 or 120 trials. Although we examined only old items, i.e. items that are shown at both study and test. The recognition test should have a minimum of 40 old items presented at study and test in addition to the same number of items presented at study but not test and new items presented at test but not study. Although results can vary between studies based on differences in semantic relatedness and distinctiveness within the word set (e.g. Hunt & Smith, 2014; Roediger et al., 2001), reliability should be equally high for these stimulus sets as for words that are randomly selected and semantically unrelated (as in our stimulus set). Other considerations for memory test design include blocked vs intermixed presentation of conditions and stimulus characteristics such as concreteness

and frequency. Based on our results, we suggest that while these factors are important for experimental manipulation and control within the memory test, the reliability would remain consistently high across changes. Based on the structure of a recognition test and to ensure a sufficient level of difficulty, at least the same number of new items and studied and unpresented items at test as studied and presented items at test (≥ 60) should be used (e.g. 60 new items, 60 studied and unpresented items, 60 studied and presented items). We could not demonstrate the effect of changing the proportion of new items because new items are not learned and accordingly are not presented at both study and test.

The examination of different trial numbers is novel and valuable because until now, the minimum number of items in a reliable memory test remained unclear, particularly when neuropsychological tests showed differences between patients and healthy controls with as few as ten items (e.g. Selective Reminding Test, Buschke & Fuld, 1974). Generalizability theory is a well-established but underused method for determining how well a test discriminates between individuals in the intended construct, and the method is especially valuable in identifying factors that can increase the reliability (such as trial number). The uniqueness of theory for evaluating reliability of the long-term memory test is apparent if considering that such investigation would not be possible using traditional CTT methods because they would require completion of the test using the same sets of words and in turn make the test invalid due to learning effect.

The results suggest that because recognition measurements are equally reliable in both languages, the test can be confidently used to discriminate between groups with different memory performance and explore individual differences or reveal effects of cognitive manipulations. The finding that individual differences were the greatest source of error suggests that many other factors can influence memory in addition to age, education, and NIBS. The presence of significant differences between groups and the existence of individual differences in memory that cannot be accounted for by age, education, NIBS, or a cognitive manipulation suggest that further examination of neurobiological and cognitive factors is warranted. Future studies can examine individual differences in-depth by applying within-subjects designs and measuring memory performance over time; multiple measurements are more likely to be accurate. Accordingly, future studies should examine the reliability of the memory test across time by measuring performance at three different time points and applying G theory. The current study only

examined memory performance at one time point and used samples from three different experiments to increase the sample size of the English and Russian groups. Future studies can also apply G theory when taking recognition measurements in other languages to confirm that any differences in results cannot be attributed to differences in the reliability of the test between languages. The recognition test results with the set of words cannot be generalized to every language because language is a fixed construct (i.e. limited to the languages included in the analysis), meaning that the reliability of the test is known only for Russian and English speakers. Since the reliability results only apply to English- and Russian-speaking younger adults, future studies should measure the test's reliability in other populations, including older adults and patients with memory disorders. Although these populations are more likely to be tested with neuropsychological tests, there is still a need for tests that are sensitive to memory changes, particularly deficits and decline (Busch et al., 2019; Elkana et al., 2016).

The recognition test has the advantage over recall tests of a larger number of trials (Murdock, 1968), better performance (Allan & Rugg, 1997), and decreased individual variability (Gomulicki, 1956; Mandler, 1968; Ozier, 1980). While recall tests may be more difficult and index recollection rather than familiarity, recognition tests may be more practical. The laboratory recognition test has the potential to be flexibly applied with different languages, items, and list lengths, according to the needs of the experiment and the research question. Specifically, the reliability of the recognition test has implications for measuring tDCS-induced neuroenhancement: studies have been conducted in different languages and countries, including Russian and English, and a reliable memory test can demonstrate when tDCS has led to changes in memory in cross-sectional measurements and over time. In conclusion, the widely-used method of applying the recognition test in the laboratory, with a different set of words in different languages for each experiment, may be as reliable as for recognition tests with a fixed format. Although the reliability of the test remains to be tested with other languages and populations, the reliability and flexibility as shown for Russian and English in younger adults may make it an optimal assessment for cognitive settings.

Appendix 1 Verbal stimuli for recognition test

Table 5

Table 5 English stimuli with Russian equivalents and associated frequency and number of letters

English	Frequency	Letters	Russian	Frequency	Letters
ballerina	1	9	балерина	8.1	8
coyote	1	6	шакал	6.5	5
crocodile	1	9	крокодил	10.7	8
fireman	1	7	пожарный	24	8
panther	1	7	пантера	2.1	7
pheasant	1	8	фазан	1.8	5
butterfly	2	9	бабочка	21.3	7
drummer	2	7	ударник	4.5	7
gangster	2	8	бандит	42.8	6
peacock	2	7	павлин	2.4	6
spider	2	6	паук	9.9	4
pigeon	3	6	голубь	13.7	6
rooster	3	7	петух	16	5
butler	4	6	слуга	17.8	5
gardener	4	8	садовник	5.4	8
plumber	4	7	слесарь	9.3	7
nightingale	4	11	соловей	7.9	7
crane	5	5	журавль	5.2	7
decorator	5	9	маляр	2.7	5
fisherman	5	9	рыбак	21.9	5
kitten	5	6	котенок	14.5	7
photographer	5	12	фотограф	12.7	8
lecturer	6	8	лектор	3.9	6
oyster	6	6	устрица	3.6	7
tiger	7	5	тигр	15.2	4
butcher	8	7	мясник	3.1	6
hostess	8	7	хозяйка	54.2	7
turtle	8	6	черепаха	6.3	8
athlete	9	7	атлет	2.5	5
chef	9	4	повар	13.3	5
monkey	9	6	обезьяна	17.4	8
mouse	10	5	мышь	37	4
dentist	12	7	стоматолог	2.5	10
philosopher	16	11	философ	30.6	7
priest	16	6	священник	39.7	9
nurse	17	5	медсестра	16.1	9
seal	17	4	тюлень	2.6	6
merchant	20	8	купец	12.5	5
painter	21	7	художник	142.2	8
architect	22	9	архитектор	33.3	10
sheep	23	5	овца	17.5	4
witness	28	7	свидетель	55.7	9
clerk	34	5	секретарь	77	9
baker	36	5	пекарь	0.9	6

Table 5 (continued)

English	Frequency	Letters	Russian	Frequency	Letters
soldier	39	7	солдат	142.2	6
pope	40	4	поп	13.5	3
queen	41	5	королева	22.4	8
engineer	42	8	инженер	63.1	7
lawyer	43	6	юрист	20.8	5
snake	44	5	змея	20.5	4
detective	52	9	детектив	10.7	8
poet	99	4	поэт	155.1	4
acrobat	1	7	акробат	2.5	7
dinosaur	1	8	динозавр	3.1	8
dolphin	1	7	дельфин	7.6	7
frog	1	4	лягушка	10.5	7
grizzly	1	7	медведь	33.4	7
gymnast	1	7	гимнаст	1.6	7
mosquito	1	8	комар	15.2	5
moth	1	4	моль	5.2	4
octopus	1	7	осьминог	1.7	8
parrot	1	6	попугай	10	7
pharmacist	1	10	аптекарь	1.6	8
snail	1	5	улитка	3	6
accountant	2	10	бухгалтер	15.5	9
astronaut	2	9	космонавт	14.6	9
biologist	2	9	биолог	7.4	6
cockroach	2	9	таракан	10.3	7
guitarist	2	9	гитарист	2.8	8
postman	2	7	почтальон	4.9	9
wasp	2	4	оса	3.2	3
robin	2	5	снегирь	1.1	7
clown	3	5	клоун	13.9	5
shark	3	5	акула	9.3	5
swan	3	4	лебедь	14.9	6
gypsy	4	5	цыган	16.9	5
magician	4	8	маг	14.6	3
pilgrim	4	7	паломник	4.1	8
worm	4	4	червь	9.9	5
wolf	6	4	волк	36.1	4
dictator	7	8	диктатор	6.4	8
martyr	8	6	мученик	7.8	7
pig	8	3	свинья	23.1	6
duck	9	4	утка	13.1	4
journalist	10	10	журналист	97	9
psychologist	10	12	психолог	29.3	8
fox	13	3	лиса	8.7	4
politician	13	10	политик	45.5	7
lion	17	4	лев	64.3	3
knight	18	6	рыцарь	18.9	6
actor	24	5	актер	114.4	5
pilot	44	5	пилот	26.9	5
guard	48	5	сторож	14.4	6
dog	75	3	собака	132.2	6
doctor	100	6	доктор	143.1	6

Table 5 (continued)

English	Frequency	Letters	Russian	Frequency	Letters
helicopter	1	10	вертолет	38.7	8
horoscope	1	9	гороскоп	3.5	8
jeans	1	5	брюки	37.2	5
jumper	1	6	свитер	15	6
ketchup	1	7	кетчуп	1.5	6
robot	1	5	робот	8.7	5
notebook	2	8	блокнот	12.8	7
pyramid	2	7	пирамида	16	8
quiz	2	4	викторина	1.3	9
rooftop	2	7	крыша	85	5
shampoo	2	7	шампунь	6	7
volcano	2	7	вулкан	6	6
drone	3	5	шмель	3.5	5
kettle	3	6	чайник	21.5	6
keyboard	4	8	клавиатура	4.8	10
yacht	4	5	яхта	9.5	4
witch	5	5	ведьма	12.1	6
wrap	5	4	шарф	9.6	4
fingerprint	6	11	отпечаток	13.1	9
kidney	6	6	почка	16.2	5
graveyard	7	9	могила	50.9	6
riot	7	4	бунт	11.8	4
infection	8	9	инфекция	26	8
chocolate	9	9	шоколад	10.9	7
glove	9	5	перчатка	19.8	8
zoo	9	3	зоопарк	12.4	7
thumb	10	5	палец	219.1	5
juice	11	5	сок	35.8	3
bubble	12	6	пузырь	16.1	6
breeze	14	6	ветер	140.3	5
cafeteria	15	9	столовая	36.5	8
petition	15	8	петиция	1.1	7
jeep	16	4	джип	14.7	4
taxi	16	4	такси	30.6	5
flame	17	5	огонь	139.8	5
angel	18	5	ангел	32.4	5
candle	18	6	свеча	31.6	5
essay	19	5	эссе	5	4
jungle	20	6	джунгли	8.2	7
garage	21	6	гараж	26.6	5
monument	21	8	памятник	63.8	8
skirt	21	5	юбка	26.3	4
tsunami	21	7	цунами	1	6
exhibition	22	10	выставка	89.9	8
whisky	23	6	виски	14.2	5
storm	26	5	гроза	15.7	5
festival	27	8	праздник	115	8
prize	28	5	приз	22.4	4
wound	28	5	рана	29.4	4
fence	30	5	забор	48.3	5
network	30	7	сеть	105.5	4

Table 5 (continued)

English	Frequency	Letters	Russian	Frequency	Letters
root	30	4	корень	63.7	6
gallery	31	7	галерея	23.2	7
van	32	3	кузов	22.1	5
interview	34	9	интервью	44	8
ocean	34	5	океан	49.8	5
pencil	34	6	карандаш	46.6	8
bench	35	5	скамейка	29	8
stomach	37	7	живот	65.6	5
comedy	39	6	комедия	22.7	7
avenue	46	6	проспект	30.4	8
pocket	46	6	карман	120.5	6
ring	47	4	кольцо	59.5	6
breakfast	53	9	завтрак	28.3	7
lake	54	4	озеро	54.9	5
leg	58	3	нога	459.2	4
beach	61	5	пляж	25.4	4
dream	64	5	сон	152.7	3
traffic	68	7	проезд	15.7	6
rain	70	4	дождь	83.2	5
bottle	76	6	бутылка	117.1	7
knife	76	5	нож	66.6	3
neck	81	4	шея	85.3	3
window	119	6	окно	280.8	4
hotel	126	5	гостиница	68.9	9
letter	145	6	письмо	304.3	6
market	155	6	рынок	283.3	5
mailbox	1	7	почта	33	5
marathon	1	8	марафон	3	7
zombie	1	6	зомби	1.3	5
wrinkle	2	7	морщина	12.2	7
galaxy	3	6	галактика	12.5	9
lipstick	3	8	помада	9.1	6
napkin	3	6	салфетка	11.3	8
telescope	4	9	телескоп	3.8	8
treasure	4	8	сокровище	15.5	9
tribe	4	5	племя	23.1	5
itch	5	4	зуд	3.5	3
laundry	5	7	прачечная	3.4	9
passport	6	8	паспорт	49.6	7
signature	6	9	подпись	40.5	7
towel	6	5	полотенце	20.5	9
bargain	7	7	сделка	52.7	6
blossom	7	7	цветок	92.4	6
ink	7	3	чернила	10.5	7
postcard	7	8	открытка	18.9	8
knot	8	4	узел	48.6	4
lotion	8	6	лосьон	2.6	6
poison	10	6	яд	16.5	2
cliff	11	5	утес	6.3	4
doll	11	4	кукла	25.3	5
graduation	11	10	выпуск	62.5	6

Table 5 (continued)

English	Frequency	Letters	Russian	Frequency	Letters
sleeve	11	6	рукав	42.9	5
tear	11	4	слеза	114.2	5
canyon	12	6	ущелье	11.1	6
grill	12	5	решетка	29	7
toilet	13	6	туалет	35.1	6
auditorium	14	10	аудитория	31.9	9
bush	14	4	куст	60.9	4
picnic	15	6	пикник	4.1	6
gown	16	4	наряд	19.5	5
jaw	16	3	челюсть	17.1	7
lung	16	4	легкое	19.6	6
spray	16	5	струя	20.3	5
ticket	16	6	билет	75.4	5
kiss	17	4	поцелуй	21.1	7
airport	19	7	аэропорт	37.2	8
crown	19	5	корона	9.4	6
ham	19	3	ветчина	5	7
clock	20	5	часы	72.5	4
slide	20	5	каток	7.8	5
tent	20	4	палатка	29.1	7
jail	21	4	тюрьма	75.4	6
drug	24	4	лекарство	39.6	9
cocktail	25	8	коктейль	8.6	8
weekend	27	7	выходной	36	8
wedding	32	7	свадьба	39.6	7
gift	33	4	подарок	75.4	7
bus	34	3	автобус	64.8	7
tongue	35	6	язык	324.9	4
bomb	36	4	бомба	35.8	5
questionnaire	37	13	опросник	2.3	8
reception	38	9	приемная	14.7	8
drawing	40	7	рисунок	179.2	7
weapon	42	6	оружие	141.8	6
coat	43	4	пальто	48.6	6
brush	44	5	щетка	9.9	5
brain	45	5	мозг	84.5	4
birth	66	5	рождение	98.5	8
box	70	3	коробка	49	7
shelter	70	7	укрытие	8.3	7
wine	72	4	вино	80.9	4
park	94	4	парк	69.5	4
film	96	4	фильм	196.8	5
bridge	98	6	мост	65.4	4
hospital	110	8	больница	96.6	8

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12144-022-02886-6>.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data Availability Our data is included as a supplementary file to this publication.

Declarations

Competing interest statement The authors declare no competing interests.

Ethical Approval The study was approved by the authors' institutional ethics committee.

Informed Consent All participants provided informed consent prior to their participation in this study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Allan, K., & Rugg, M. D. (1997). An event-related potential study of explicit memory on tests of cued recall and recognition. *Neuropsychologia*, 35, 387–397.
- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole.
- Atilgan, H. (2013). Sample Size for Estimation of G and Phi Coefficients in Generalizability Theory. *Eurasian Journal of Educational Research*, 51, 215–227.
- Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher*, 34, 960–992.
- Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1–21.
- Busch, R. M., Hogue, O., Ferguson, L., Parsons, M. W., Kubu, C. S., & Floden, D. P. (2019). Validation of computerized episodic memory measures in a diverse clinical sample referred for neuropsychological assessment. *The Clinical Neuropsychologist*, 33, 557–570.
- Buschke, H., & Fuld, P. A. (1974). Evaluating storage, retention, and retrieval in disordered memory and learning. *Neurology*, 24, 1019–1025.
- Cardinet, J., Johnson, S., & Pini, G. (2010). *Quantitative methods series*. Applying generalizability theory using EduG. Routledge/Taylor & Francis Group.
- Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 13, 119–135.
- Cheke, L. G., & Clayton, N. S. (2015). The six blind men and the elephant: Are episodic memory tasks tests of different things or different tests of the same thing? *Journal of Experimental Child Psychology*, 137, 164–171.
- Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33, 497–505.
- Cronbach, L. J. (1947). Test “reliability”: Its meaning and determination. *Psychometrika*, 12, 1–16.
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16, 137–163.
- Elkana, O., Eisikovits, O. R., Oren, N., Betzale, V., Giladi, N., & Ash, E. L. (2016). Sensitivity of neuropsychological tests to identify cognitive decline in highly educated elderly individuals: 12 months follow up. *Journal of Alzheimer's Disease*, 49, 607–616.
- Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1, 103–111.
- Gomulicki, B. R. (1956). Individual differences in recall. *Journal of Personality*, 24, 387–400.
- Hasselmo, M. E. (2011). *How we remember: Brain mechanisms of episodic memory*. MIT press.
- Hunt, R. R., & Smith, R. E. (2014). How distinctive processing enhances hits and reduces false alarms. *Journal of Memory and Language*, 75, 45–57.
- Larrabee, G. J., & Curtiss, G. (1995). Construct validity of various verbal and visual memory tests. *Journal of Clinical and Experimental Neuropsychology*, 17, 536–547.
- Livingston, S. A. (1972). Criterion-referenced applications of classical test theory 1, 2. *Journal of Educational Measurement*, 9, 13–26.
- Mandler, G. (1968). Organized recall: Individual functions. *Psychonomic Science*, 13, 235–236.
- Medvedev, O. N., Krägeloh, C. U., Narayanan, A., & Siegert, R. J. (2017). Measuring mindfulness: Applying generalizability theory to distinguish between state and trait. *Mindfulness*, 8, 1036–1046.
- Medvedeva, A. (2019). *NIBS as a research tool in studying and enhancing episodic memory in the left prefrontal cortex* (Doctoral dissertation). Kingston University.
- Medvedeva, A., Materassi, M., Neacsu, V., Beresford-Webb, J., Hussin, A., Khan, N., ... & Galli, G. (2019). Effects of anodal transcranial direct current stimulation over the ventrolateral prefrontal cortex on episodic memory formation and retrieval. *Cerebral Cortex*, 29, 657–665.
- Melcher, D., & Murphy, B. (2011). The role of semantic interference in limiting memory for the details of visual scenes. *Frontiers in Psychology*, 2, 262.
- Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Journal Structural Equation Modeling: A Multidisciplinary Journal*, 2, 255–273.
- Murdock, B. B., Jr. (1968). Serial order effects in short-term memory. *Journal of Experimental Psychology*, 76, 1–15.
- Otani, H., & Schwartz, B. L. (Eds.). (2019). *Handbook of Research Methods in Human Memory*. Routledge.
- Ozier, M. (1980). Individual differences in free recall: When some people remember better than others. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 14 (pp. 309–364). New York: Academic Press.
- Petrovskaya, A., Kirillov, B., Asmolova, A., Galli, G., Feurra, M., & Medvedeva, A. (2020). Examining the effects of transcranial direct current stimulation on human episodic memory with machine learning. *Plos One*, 15, e0235179.
- Rahimi-Golkhandan, S., Maruff, P., Darby, D., & Wilson, P. (2012). Barriers to repeated assessment of verbal learning and memory:

- A comparison of international shopping list task and rey auditory verbal learning test on build-up of proactive interference. *Archives of Clinical Neuropsychology*, 27, 790–795.
- Renoult, L., & Rugg, M. (2019, October 9). An historical perspective on Endel Tulving's episodic-semantic dichotomy. Preprint. <https://doi.org/10.31234/osf.io/uzdpf>
- Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, 8, 385–407.
- Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, 53, 1–25.
- Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin*, 145, 79–139. <https://doi.org/10.1037/bul0000176>

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