PLOS ONE

Measuring individual differences in cognitive abilities in the lab and on the web --Manuscript Draft--

Manuscript Number:	PONE-D-19-22438R1			
Article Type:	Research Article			
Full Title:	Measuring individual differences in cognitive abilities in the lab and on the web			
Short Title:	Measuring individual differences on the web			
Corresponding Author:	Simon Ruiz Eberhard Karls Universitat Tubingen Tübingen, GERMANY			
Keywords:	web-based testing; measurement equivalence; cognitive individual differences; Working Memory; declarative memory			
Abstract:	The present study compared lab-based and web-based versions of cognitive individual difference measures widely used in second language research (working memory and declarative memory). Our objective was to validate web-based versions of these tests for future research and to make these measures available for the wider second language research community, thus contributing to the study of individual differences in language learning. The establishment of measurement equivalence of the two administration modes is important because web-based testing allows researchers to address methodological challenges such as restricted population sampling, low statistical power, and small sample sizes. Our results indicate that the lab-based and web-based versions of the tests were equivalent, i.e., scores of the two test modes correlated. The strength of the relationships, however, varied as a function of the kind of measure, with equivalence appearing to be stronger in both the working memory and the verbal declarative memory tests, and less so in the nonverbal declarative memory test. Overall, the study provides evidence that web-based testing of cognitive abilities can produce similar performance scores as in the lab.			
Order of Authors:	Simon Ruiz			
	Xiaobin Chen			
	Patrick Rebuschat			
	Detmar Meurers			
Opposed Reviewers:				
Response to Reviewers:	 2nd November 2019 Thank you for the specific feedback on the manuscript entitled "Measuring individual differences in cognitive abilities in the lab and on the web". Here is our response on how we took the feedback into account in revising the paper: Reviewer #2: Thank you for the opportunity to review this paper. It is an interesting study that compares lab-based and web-based versions of memory tests in a sample of adults with the aim of validating the web-based version. The article is well-written and the study is set up well in general. I listed a few specific comments below: *P3, I.59: when referring to the benefits of web-based testing it would be interesting to refer to other possible simultaneous testing strategies available. For instance, there are many tests that can be answered by individuals in school or university settings that might have similar benefits compared to web-based versions, so it would be important to emphasize what is the specific advantage of this type of tool. While the established paper-and-pencil tests naturally can be administered by 			

	time that could be used for teaching and learning activities. Conducting such paper-
	and-pencil tests in class would also be more of an issue in school cultures in which standardized testing is less common than in the US. We added a new paragraph that discusses other methodological advantages of (remote) web-based testing in comparison to other forms of simultaneous delivery of tests, such as traditional paper-pencil and (offline) computer- based testing (p. 3).
	*P4, I.75: please provide an argument of why are you only looking at one type of equivalence.
	The following argument was added (p. 5):
	Considering that this study is a subcomponent of the dissertation research of the first author, limiting funding and time (see limitations below), we focused the investigation on one type of measurement equivalence, the first type: Do people who have relatively high values in one of tests also have relatively high values on the other test, and the other way around?
	*P5, I.92: throughout the paper there are several mentions to L2 research, however the issue of small sample size and low power are not restricted to that research area. I would expand the claim to many other situations where methodological issues related to testing are a challenge.
	The discussion of the methodological issues was expanded, including reference to low statistical power and small sample sizes being problematic in other research fields and the ongoing debate in the so-called replication crisis in psychology (p. 5-6).
	*P8, I.165: the fact that the sample was not full due to technical reasons requires more explanation. Is this related to possible flaws of web-based testing? If so, it should be included in the discussion.
	We added the following explanation (p. 9):
	Additionally, participant numbers differed across test versions due to technical difficulties (i.e., participants erroneously entered their responses using the keyboard [Web-based CVMT]; and data was missing for one participant [Web-based MLAT5]; see description and Table 1 below, and Discussion).
	and a discussion of these technical shortcomings is included in the Discussion section (p. 18).
	*Dicussion: I think it would be important to discuss the limitations of the study and also of the findings.
	We added limitations of the study and findings in the Discussion section (p. 18).
	Yours sincerely,
	Simón Ruiz, Xiaobin Chen, Patrick Rebuschat, and Detmar Meurers
Additional Information:	
Question	Response
Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review	Our research was supported by the LEAD Graduate School and Research Network (grant DFG-GSC1028), a project of the Excellence Initiative of the German federal and state governments. We also acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of University of Tübingen The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

requirements. View published research articles from <u>PLOS ONE</u> for specific examples.

This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate.

Unfunded studies

Enter: The author(s) received no specific funding for this work.

Funded studies

Enter a statement with the following details:

- Initials of the authors who received each
 award
- Grant numbers awarded to each author
- The full name of each funder
- URL of each funder website
- Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript?
- NO Include this sentence at the end of your statement: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
- YES Specify the role(s) played.

* typeset

* typeset	
Competing Interests	The authors have declared that no competing interests exist.
Use the instructions below to enter a	
competing interest statement for this	
submission. On behalf of all authors,	
disclose any competing interests that	
could be perceived to bias this	
work—acknowledging all financial support	
and any other relevant financial or non-	
financial competing interests.	
This statement will appear in the	
published article if the submission is	
accepted. Please make sure it is	
accurate. View published research articles	
from <u>PLOS ONE</u> for specific examples.	

NO authors have competing interests	
Enter: The authors have declared that no competing interests exist.	
Authors with competing interests	
Enter competing interest details beginning with this statement:	
I have read the journal's policy and the authors of this manuscript have the following competing interests: [insert competing interests here]	
* typeset	
Ethics Statement Enter an ethics statement for this submission. This statement is required if	This research was approved by the Commission for Ethics in Psychological Research, University of Tübingen, and all participants provided written informed consent prior to commencement of the study.
the study involved:	
Human participantsHuman specimens or tissue	
Vertebrate animals or cephalopods	
Vertebrate embryos or tissuesField research	
Write "N/A" if the submission does not	
require an ethics statement.	
General guidance is provided below.	
Consult the submission guidelines for	
detailed instructions. Make sure that all	
information entered here is included in the Methods section of the manuscript.	

Format for specific study types

Human Subject Research (involving human participants and/or tissue)

- Give the name of the institutional review board or ethics committee that approved the study
- Include the approval number and/or a statement indicating approval of this research
- Indicate the form of consent obtained (written/oral) or the reason that consent was not obtained (e.g. the data were analyzed anonymously)

Animal Research (involving vertebrate

animals, embryos or tissues)

- Provide the name of the Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board that reviewed the study protocol, and indicate whether they approved this research or granted a formal waiver of ethical approval
- Include an approval number if one was obtained
- If the study involved non-human primates, add additional details about animal welfare and steps taken to ameliorate suffering
- If anesthesia, euthanasia, or any kind of animal sacrifice is part of the study, include briefly which substances and/or methods were applied

Field Research

Include the following details if this study involves the collection of plant, animal, or other materials from a natural setting:

- Field permit number
- Name of the institution or relevant body that granted permission

Data Availability

Authors are required to make all data underlying the findings described fully available, without restriction, and from the time of publication. PLOS allows rare exceptions to address legal and ethical concerns. See the <u>PLOS Data Policy</u> and FAQ for detailed information.

Yes - all data are fully available without restriction

A Data Availability Statement describing where the data can be found is required at submission. Your answers to this question constitute the Data Availability Statement and will be published in the article , if accepted. Important: Stating 'data available on request from the author' is not sufficient. If your data are only available upon request, select 'No' for the first question and explain your exceptional situation in the text box. Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
 Describe where the data may be found in full sentences. If you are copying our sample text, replace any instances of XXX with the appropriate details. If the data are held or will be held in a public repository, include URLs, accession numbers or DOIs. If this information will only be available after acceptance, indicate this by ticking the box below. For example: <i>All XXX files are available from the XXX database (accession number(s) XXX, XXX.)</i>. If the data are all contained within the manuscript and/or Supporting Information files, enter the following: <i>All relevant data are within the manuscript and its Supporting Information files.</i> If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so. For example: 	
researchers who meet the criteria for access to confidential data. The data underlying the results presented in the study are available from (include the name of the third party	

 and contact information or URL). This text is appropriate if the data are owned by a third party and authors do not have permission to share the data. * typeset 	
Additional data availability information:	



8th August 2019

We would like our manuscript entitled "**Measuring individual differences in cognitive abilities in the lab and on the web**" to be considered for publication in *PLOS ONE*.

Our manuscript reports the results of a study that compared lab-based and web-based versions of individual difference measures that are widely investigated in language learning research (working memory and declarative memory). Our objective was to validate web-based versions of these cognitive tests for future research and to make these measures freely available for the wider community, thus contributing to the study of individual differences in language learning. The establishment of measurement equivalence of the two administration modes is important because web-based testing allows researchers to address methodological challenges such as restricted population sampling, low statistical power, and small sample sizes.

Our results indicate that the lab-based and web-based versions of the tests were equivalent, i.e. scores of the two test modes correlated. The strength of the relationships, however, varied as a function of the kind of measure, with equivalence appearing to be stronger in both the working memory and the verbal declarative memory tests, and less so in the nonverbal declarative memory test. Overall, the study provides evidence that web-based testing of cognitive abilities can produce similar performance scores as in the lab.

We believe that this manuscript is particularly suitable for the audience of *PLOS ONE* because it concerns measuring cognitive abilities on the web, which could be a feasible alternative to tackle some of the current methodological issues found in language learning research conducted in lab-based settings.

We suggest Michael Thomas Ullman be an Academic Editor for this work.

The data presented here has not previously been published, and has not been submitted for publication to another journal.

Many thanks in advance for considering our manuscript for your journal.

Yours sincerely,

Simón Ruiz, Xiaobin Chen, Patrick Rebuschat, and Detmar Meurers

<u>±</u>

1

1	Measuring individual differences in cognitive abilities in the lab and on the web
2	
3	
4	Simón Ruiz ^{1*} , Xiaobin Chen ² , Patrick Rebuschat ^{1,3} Detmar Meurers ^{1,4}
5	
6	
7 8	
9	¹ [¶] LEAD Graduate School and Research Network, University of Tübingen, Tübingen, Germany
10	ELTE Graduate School and Research Petwork, Oniversity of Paoingen, Paoingen, Germany
11	² [¶] Department of Theoretical and Applied Linguistics, University of Cambridge, United Kingdom
12	
13	³ [¶] Department of Linguistics and English Language, Lancaster University, Lancaster, United
14	Kingdom
15	
16	⁴ Department of Linguistics, University of Tübingen, Tübingen, Germany
17	
18	* Corresponding author
19	E-mail: simon.ruiz-hernandez@sfs.uni-tuebingen.de (SR)
20	
21 22	[¶] These authors contributed equally to this work.
22	These autions contributed equally to this work.
23 24	

25 Abstract

26 The present study compared lab-based and web-based versions of cognitive individual difference measures widely used in second language research (working memory and declarative 27 28 memory). Our objective was to validate web-based versions of these tests for future research and 29 to make these measures available for the wider second language research community, thus 30 contributing to the study of individual differences in language learning. The establishment of 31 measurement equivalence of the two administration modes is important because web-based testing 32 allows researchers to address methodological challenges such as restricted population sampling, 33 low statistical power, and small sample sizes. Our results indicate that the lab-based and web-34 based versions of the tests were equivalent, i.e., scores of the two test modes correlated. The 35 strength of the relationships, however, varied as a function of the kind of measure, with 36 equivalence appearing to be stronger in both the working memory and the verbal declarative 37 memory tests, and less so in the nonverbal declarative memory test. Overall, the study provides 38 evidence that web-based testing of cognitive abilities can produce similar performance scores as 39 in the lab.

40

41 Introduction

Individual differences can greatly affect how we acquire and process language [1-3] and mediate and/moderate the effectiveness of instruction [4]. In adult language learning, for example, learners' cognitive abilities have great explanatory power in accounting for differences in learning outcomes ([5-6]). Among these, working memory and declarative memory are considered to be particularly important sources of learner variation (e.g., [7-10]; see [4, 11], for reviews). 47 The effect of working memory and declarative memory on language learning has been 48 primarily studied in lab settings, i.e., in well-controlled environments where participants are tested 49 individually. While this choice is methodologically sound, it can also negatively affect sample size 50 and population sampling [13, 14]. Lab-based testing generally means testing participants 51 individually and sequentially, which is labor-intensive and could explain why lab studies tend to 52 have (too) few participants to allow for meaningful generalization. For example, Plonsky [13] 53 found that the typical sample size in L2 studies was 19 participants, and Lindstromberg [15] 54 recently reported a similar small average sample size of 20 participants. Moreover, many (if not 55 most) lab studies in L2 research draw their sample from the surrounding student population, which 56 is understandable given the ease of access, but also means that samples are often not representative 57 of the population of interest.

58 Conducting second language research by means of remote testing via the web could 59 alleviate some of these concerns. For example, web-based testing facilitates the acquisition of large 60 amounts of data since participants can be tested simultaneously, and test administration can also 61 be more cost-effective than research conducted in the lab [15]. Importantly, web-based 62 experimenting has been found to be a reliable and effective research tool [16,17, 18].

The present study compared lab-based and web-based versions of cognitive tests that are widely used in second language research. The intent was to compare performance of measures as they are originally used in the lab with their corresponding online versions. In doing so, our objective was to validate the web-based tests for use in subsequent research and to make these available to the wider second language research community. The sharing of tasks, especially of tasks that permit the collection of substantial amounts of data via the web, will be an important component in reducing the data problem in SLA. Making these specific tasks available will also 70 contribute directly to our understanding of individual differences in L2 acquisition. To support 71 such task sharing and use, it is essential to first establish the validity of the online versions of the 72 tasks (on a par with what is established about the offline versions). With this in mind, the study set 73 out to establish measurement equivalence between lab-based and web-based tests of working 74 memory and declarative memory.

75 According to Gwaltney, Shields and Shiffman ([19], p. 323), measurement equivalence can 76 be established if "1) the rank orders of scores of individuals tested in alternative modes closely 77 approximate each other; and 2) the means, dispersions, and shapes of the score distributions are 78 approximately the same". The first type of equivalence is related to whether differences found in 79 one measurement are also systematically found in the other. This means that, although the two 80 measurements estimate two different numbers, these numbers have a systematic and very clear 81 relationship to each other. The second type concerns whether two measurements yield the same 82 numbers. Here, we focus on the former type of equivalence. More specifically, we compare the 83 differential performance generated by two versions of tests measuring working memory and 84 declarative memory capacities in lab-based and web-based settings, with the aim to determine 85 whether the two versions are equivalent with respect to the relationships between scores. 86 Establishing measurement equivalence between these two administration modes is essential for 87 several reasons. First, it is necessary to show that the results of web-based studies are comparable 88 to those of previous research, which have predominantly obtained from data gathered in lab-based 89 settings. Second, it is imperative to ensure that cognitive constructs are measured in the same way 90 in both test modes. Finally, it is important to ascertain whether lab-based and web-based measures 91 are equivalent because, if they are, web-based testing could be a feasible alternative to address

some of the current methodological issues found in L2 research conducted in lab-based settings,
such as underpowered studies and small sample sizes, among others [13, 14].

94

95 Working memory

Working memory refers to the capacity to simultaneously process and retain information while carrying out complex cognitive tasks such as language learning, comprehension and production [20]. Following Baddeley and colleagues (e.g., [21]), working memory is a multicomponent system that consists of storage subsystems that are responsible for holding visualspatial and auditory information, an episodic buffer that acts as a link between the storage subsystems and long-term memory, and a central executive that functions as an attentional control system.

103 In L2 learning, working memory appears to assist learners to jointly process form, meaning 104 and use of language forms at the same time. More specifically, working memory is involved in 105 key cognitive processes such as decision making, attention control, explicit deduction, information 106 retrieval and analogical reasoning [4]. Moreover, working memory is also important for retaining 107 metalinguistic information while comprehending and producing L2 language [22]. In this regard, 108 meta-analytic work has reported the important role of working memory in L2 comprehension and 109 production (e.g., [23-25]). For example, Linck et al. ([25], p. 873) found that working memory has a positive impact on L2 comprehension outcomes (r = 0.24). Likewise, Jeon and Yamashita's [24] 110 111 meta-analysis also showed that working memory is related to L2 reading comprehension (r = 0.42). 112 Regarding production, meta-analytic research has, too, indicated a significant association with 113 working memory (e.g., [25]). In this case, Linck et al. ([25], p. 873) found a positive correlation 114 for productive outcomes as well (r = 0.27).

115 Working memory is often measured by means of simple or complex span tasks. Simple 116 span tasks (e.g., digit span and letter span) involve recalling short lists of items, and they seek to 117 gauge the storage aspect of working memory [26]. Complex span tasks, such as the operation span 118 task (OSpan; [27]), on the other hand, entail remembering stimuli while performing a secondary 119 task, and are thought to tax both processing (attention) and storage (memory) components of 120 working memory [21]. Here, we focus on a complex task, namely the OSpan. This complex task 121 has been found to be a valid and reliable measure of working memory capacity [28], and has also 122 been recommended as a more accurate measure to examine the association between working 123 memory and L2 processing and learning [29].

124

125 **Declarative memory**

Declarative memory is the capacity to consciously recall and use information [30]. The declarative memory system is one of the long-term memory systems in the brain [31]. It is mainly responsible for the processing, storage, and retrieval of information about facts (semantic knowledge) and events (episodic knowledge; [32, 33]). Learning in the declarative memory system is quick, intentional, and attention-driven [34].

Substantial research has now investigated the role of declarative memory in first and second language acquisition [35]. In first language acquisition, declarative memory appears to be involved in the processing, storage and learning of both arbitrary linguistic knowledge (e.g., word meanings) as well as rule-governed aspects of language (e.g., generalizing grammar rules [36,37]). In the case of L2 acquisition, declarative memory appears to underpin the learning, storage and processing of L2 vocabulary and grammar [36,37], at least in the earliest phases of acquisition [35, 137 38]. Several studies (e.g., [2, 9, 38, 39]) has confirmed the predictive ability of declarative memory
138 to explain variation in L2 attainment.

Declarative memory has been tested through recall and recognition tasks (e.g., 38, 39), both
verbal, such as the paired associates subtest of the Modern Language Aptitude Test (MLAT5;
[40]), and nonverbal, such as the Continuous Visual Memory Task (CVMT; [41]).

142

143 **The present study**

144 The main goal of the present study was to provide web-based versions of commonly employed 145 individual difference measures in second language research, in order to make them usable in large-146 scale intervention studies (generally in authentic, real-life learning contexts). To that end, we 147 examined whether lab-based and web-based versions of working memory and declarative memory 148 tests yield similar performance scores, i.e., whether the two versions were equivalent or 149 comparable. More specifically, we assessed whether the values of one type of mode of 150 administration corresponded to the values in the other mode (i.e., first type of equivalence). In 151 other words, are the differences in scores constant, or parallel in the two ways of measuring? The 152 web-based versions are freely available; to use the test, please send an email to the first author.

153

154 Methods

155 **Ethics statement**

This research was approved by the Commission for Ethics in Psychological Research, University of Tübingen, and all participants provided written informed consent prior to commencement of the study.

159

160 **Participants**

161 Fifty participants (37 women and 13 men), with a mean age of 26.4 years (SD = 4.2), took 162 part in the study. Most participants were native speakers of German (72%), followed by Russian 163 (8%), Spanish (6%), Chinese (4%), English, Hungarian, Persian, Serbian and Vietnamese (2%) 164 each). Seven (14%) participants did not complete the second half of the study (i.e., web-based 165 testing). Additionally, participant numbers differed across test versions due to technical difficulties 166 (see Results; Table 1). Twenty-seven participants were graduate students (54%), and twenty-three 167 were undergraduates (46%). Participants self-reported English proficiency, with most being 168 advanced learners (82%), followed by intermediate (18%). All subjects gave informed consent and 169 received €20 for participating.

170

171 Materials

172 Three cognitive tests were administered, one assessing working memory capacity, and 173 two indexing verbal and nonverbal declarative memory capacity, respectively. In the lab-based 174 context, working memory and nonverbal declarative memory tests were programmed and 175 delivered via E-Prime v2.0 [42]; the verbal declarative memory test was applied in paper-pencil 176 form, as originally developed and delivered. For the web-based mode, versions of the three 177 cognitive tests were developed for this study using Java with the GoogleWeb Toolkit 178 (http://www.gwtproject.org), and were accessible from all browsers. The tests are described 179 below.

180

Working memory. To assess participants' working memory capacity, an adapted version of the
Automated Operation Span Task (OSpan; [43]), a computerized form of the complex span task

183 created by Turner and Engle [27], was used [9, 17]. This adaptation was based on the Klingon 184 Span Task developed by Hicks et al. [17], and consisted of replacing letters (the original stimuli 185 to be remembered in the OSpan task) with Klingon symbols. Hicks et al. implemented this 186 change because their research showed that participants were cheating by writing down the letter 187 memoranda in the web-based version of the classic OSpan.

188 The task took approximately 25 minutes to complete, and was divided into a practice phase 189 and a testing phase. In the practice phase, participants were first presented with a series of Klingon 190 symbols on the screen, and were asked to remember them in the order they had appeared at the 191 end of each trial (i.e., symbol recall). Next, participants were asked to solve a series of simple math 192 operations (e.g., 5 * 2 + 1 = ?). Finally, subjects performed the symbol recall while also solving the 193 math problems, as they would do later in the actual testing phase. After the practice phase, 194 participants were presented with the real trials, which consisted of a list of 15 sets of 3-7 195 randomized symbols that appeared intermixed with the equations, totaling 75 symbols and 75 math 196 problems. At the end of each set, participants were asked to recall the symbols in the sequence 197 they had been shown. An individual time limit to answer the math problems in the real trials was 198 derived from the average response time plus 2.5 standard deviations taken during the math practice 199 section. Following Unsworth et al. [46], a partial score (i.e., total number of correct symbols 200 recalled in the correct order) was taken as the OSpan score (see [28], for a description of scoring 201 procedures). The highest possible score was 75.

202

203 Verbal declarative memory. The Modern Language Aptitude Test, Part 5, Paired

Associates (MLAT5; [40]), was used as a verbal measure of declarative memory [9, 38, 39]. The

205 MLAT5 required participants to memorize artificial, pseudo-Kurdish words and their meanings

in English. Participants first studied 24-word association pairs for two minutes, and then
completed a two-minute practice section. During the practice section, the list of foreign words
and their English equivalents were made available for participants to refer back if they needed to.
Finally, subjects completed a timed multiple-choice test (four minutes), in which they were
asked to select the English meaning of each of the 24 pseudo-Kurdish words from five options
previously seen at the memorization stage. For each correct response, one point was awarded,
yielding a total score of 24 points. The test duration was 8 minutes.

213

214 **Nonverbal declarative memory.** The Continuous Visual Memory Task (CVMT; [46]) 215 was included as an assessment of nonverbal declarative memory [9, 38, 39]. The CVMT is a 216 visual recognition test that involves asking participants to first view a collection of complex 217 abstract designs on the screen, and then to indicate whether the image they just saw was novel 218 ("new") in the collection, or they had seen the image before ("old"). Seven of the designs were 219 "old" (target items), and 63 were "new" (distractors). Throughout the task, the target items 220 appeared seven times (49 trials), and the distractors only once (63 trials). All items were 221 presented in a random but fixed order, each one appearing for two seconds. After the two 222 seconds, participants were instructed to respond to the "OLD or NEW?" prompt on the screen. In the lab-based setting, subjects indicated their choice by mouse clicking either left for "NEW", or 223 224 right for "OLD". In the web-based setting, they responded by pressing either the "N" key for 225 "NEW", or the "O" key for "OLD" on the keyboard. Overall, the CVMT required 10 minutes to 226 be completed. For each participant, a d'(d-prime) score [44] for CVMT was computed. The d' 227 score was used to account for the possible participants' response bias toward choosing "OLD" or 228 "NEW".

229

230 **Procedure**

As previously noted, participants completed two cognitive testing sessions, one in the lab and one on the web. In the lab-based session, in the presence of a proctor, each subject was tested individually. After providing informed consent, participants took the three cognitive tests under investigation in fixed order: OSpan, CVMT, and MLAT5. They were then asked to fill in a background questionnaire. The whole lab-based session took about 40 minutes.

236 For the web-based session, each subject was sent an email containing a unique web link with a 237 personalized code, that when clicked, took them to an interface housing the web-based versions of 238 the cognitive tests. To prevent participants from taking the tests multiple times, the link became 239 nonfunctional once they had submitted their responses in the last test (i.e., MLAT5). In the email, 240 participants were also informed that the web-based session lasted about 40 minutes, and had to be 241 completed within a week. On the interface, following informed consent, subjects were given 242 general instructions in accordance with the web-based nature of the experiment. These instructions 243 included completing the experiment in a quiet place without interruption, and from start to finish 244 in one sitting. Participants were also instructed not to use the browser's back button, or refresh the 245 browser page, or close the browser window. Importantly, they were told not to take any notes 246 during the entire experiment. The tests were taken in the same fixed order as in the lab-based 247 session. The mean period between the first and second testing was 45.7 days (SD = 4.1).

248

249 **Results**

All data were analyzed using the statistical software package R version 3.3.2 (R Core
Team, 2016). Missing data was ignored (complete case analysis). From a temporal point of view,

252	lab scores were used to predict web scores in the linear regression models. To verify normality,
253	model residuals were visually inspected. Reliability was assessed using Cronbach's alpha.
254	Following Kane et al. [45], for the lab-based working memory test (OSpan-Lab-based),
255	reliability was assessed by calculating the proportion of correctly recalled Klingon symbols per
256	each of the 15 trials in the test (e.g., one out of four symbols correctly recalled corresponded to a
257	proportion of .25). For the web-based working memory test (OSpan-Web-based), however,
258	internal consistency is not reported, since it was not technically possible to perform a detailed
259	item-based analysis. Descriptive statistics are presented first, followed by correlations, internal
260	consistency estimates (Cronbach's alpha), and the results of linear regression analyses.

261

262 **Descriptive statistics**

Table 1 presents the descriptive statistics summarizing participants' performance on thethree cognitive tests under investigation in both test modes.

265

266	Table 1. Descriptive sta	atistics for comparison	of lab-based and web-based testing.	
	····· · · · · · · · · · · · · · · · ·	· · · · · · · · · · · · · · · · · · ·		

Test	N	М	SD	Skew	Kurtosis
OSpan Lab-based	50	25.78	13.34	0.61	2.90
OSpan Web-based	43	29.79	15.42	0.67	3.26
MLAT5 Lab-based	50	17.92	5.50	-0.64	2.49
MLAT5 Web-based	42	19.10	5.81	-1.19	3.58
CVMT Lab-based	49	1.99	0.46	0.23	3.35
CVMT Web-based	40	2.30	0.63	0.73	3.32

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task.

267

268 Correlations

- Table 2 and Fig 1 show the correlations between/among the different versions of the
- 270 individual difference tests.
- 271

272 Table 2. Correlations between lab-based and web-based scores for individual difference

273 **tests.**

Test	OSpan	OSpan	MLAT5	MLAT5	CVMT
	Lab-based	Web-based	Lab-based	Web-based	Lab-based
OSpan Web-based	.80				
MLAT5 Lab-based	.40	.51			
MLAT5 Web-based	.32	.40	.82		
CVMT Lab-based	.19	.31	.42	.23	
CVMT Web-based	.21	.30	.21	.19	.55

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task.

274

Fig 1. Scatterplots of the correlation of each pair of lab-based and web-based versions of

276 individual difference measures.

14

- 277 OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern
- 278 Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous
- 279 Visual Memory Task.
- 280
- 281 Reliability
- Table 3 presents Cronbach's alpha values of individual test versions.
- 283

Table 3. Cronbach's alphas for cognitive test versions.

Test	Cronbach's alpha
OSpan Lab-based	.86
MLAT5 Lab-based	.77
MLAT5 Web-based	.93
CVMT Lab-based	.63
CVMT Web-based	.67

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task.

285

286 Regression analysis

287 The results of the regression analyses are displayed in Table 4. For the working memory 288 test (OSpan), the unstandardized coefficient was .89 (β = .77, *SE* = 0.10, *p* < .001). For the verbal 289 declarative memory test (MLAT5), the unstandardized coefficient was .83 (β = .78, *SE* = 0.09, *p*

290 <.001). And for the nonverbal declarative memory test (CVMT), the unstandardized coefficient

291 was .74 (β = .54, *SE* = 0.19, *p* < .001). Overall, the results indicated that the lab-based and web-292 based scores are substantially related.

293

Table 4. Regression for comparison of lab-based and web-based scores.

Test	Unstandardized coefficient ^a	SE	р
OSpan	0.89 (.77)	0.10	< .001
MLAT5	0.83 (.78)	0.09	< .001
CVMT	0.74 (.54)	0.19	< .001

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task. ^aThe standardized coefficient (β) in parentheses.

295

296 **Discussion**

297 Studies on individual differences in language learning frequently assess the working 298 memory and declarative memory capacities of their participants in order to determine the effect 299 of these cognitive variables on learning outcomes. Most of this research, however, is conducted 300 in lab-based settings, which often implies relatively small sample size and a restricted population 301 sampling. Both of these methodological challenges can be addressed by means of remote testing 302 via the web. In the present study, we compared lab-based and web-based individual difference 303 measures in order to validate web-based tests for future research. The type of comparison 304 contributes significantly to ongoing efforts to improve the methodological robustness of current 305 second language research [47]. If web-based testing can be shown to yield comparable results to

306 lab-based testing, researchers will be able to reach more participants for their studies, which, in 307 turn, can help alleviate some of the current concerns in L2 research (e.g., low statistical power, 308 non-representative population samples, and small sample sizes). In addition, demonstrating the 309 equivalence of lab-based and web-based measures of the same individual difference constructs is 310 essential for the comparability of results across studies. Crucially, establishing measurement 311 equivalence between lab-based and web-based versions will also provide assurance that the tests 312 are measuring cognitive constructs the same way regardless of administration mode [16, 48].

313 The results indicated that the scores in the lab-based and web-based versions of three 314 cognitive tests (MLAT5, CVMT, OSpan) were equivalent in the sense that differences in 315 performance were constant in the two versions. This suggests that participants who had relatively 316 high values in one task also had relatively high values in the second, or the other way around. 317 However, the strength of the association depended on the test. In both the working memory test 318 (OSpan) and the verbal declarative memory test (MLAT5) the scores were more strongly 319 correlated ($\beta = .77$ and $\beta = .78$, respectively); for the nonverbal declarative test (CVMT), 320 equivalence appears to be weaker ($\beta = .54$). On the whole, the correlations reported here between 321 lab-based and web-based scores are consistent with the assumption that both versions seem to 322 likely measure the same cognitive construct, at least for the working memory test (OSpan) and 323 the verbal declarative memory test (MLAT5), and, to a lesser extent, for the nonverbal 324 declarative test (CVMT).

A possible explanation for the weaker equivalence found for the versions of the nonverbal declarative test (CVMT) is perhaps the difference in the way responses to the visual stimuli were input in the two testing modes. Recall that in the lab-based version, participants 328 used left ("NEW") or right ("OLD") mouse clicking to enter their response, whereas in the web-329 based version, they used the keyboard ("N" and "O" keys). This modification was made to the 330 web-based version because of technical reasons, i.e., the browser window may not register the 331 participants' response if the cursor is not over a certain area on the page, which in itself may 332 cause problems of missing data. It has been previously reported that participants in web-based 333 research are prone to make errors when using the keyboard to enter their responses [49], which 334 in this case might have affected the results of the comparison between lab-based and web-based 335 versions of CVMT. Further studies comparing performance between the two versions may 336 benefit from gathering data via touch input instead, which might overcome the technical 337 difficulty of employing mouse clicking for web-based data collection reported here.

338

339 Conclusion

340 This study aimed to establish the validity of using web-based versions of established 341 offline tasks. As such, the study has provided evidence that it is possible to measure individual 342 differences in cognitive abilities on the web and obtain similar performance as in the lab. The 343 lab-based and web-based versions of the three cognitive tests are comparable or equivalent. 344 However, given that they do not perfectly correlate, we recommend using one of the two modes 345 within one study and not comparing individual scores from one mode with scores from the other. 346 Moreover, the extent to which the measures are equivalent varies according to the test. In this 347 sense, we are confident that the two versions for the working memory test (OSpan) and the 348 verbal declarative memory (MLAT5) are fairly possibly measuring the same construct, but we 349 refrain from making such a strong statement for the nonverbal declarative test (CVMT), where

350	the two modes might still plausibly measure strongly different aspects as well. Our research has
351	shown that collecting experimentally controlled data on cognitive individual differences typically
352	used in L2 research in the Internet is feasible and comparable to lab-based collection.
353	Consequently, some of these web-based versions could very well be incorporated, for example,
354	in future web-based intervention studies on second language learning, thereby contributing to the
355	scaling up of data collection in the field [50-52].
356 357 358 359	Acknowledgements We would like to thank Johann Jacoby, for his invaluable for his valuable suggestions that
360	strengthened our experimental design and analysis.
361	
362	References
363	1. Kidd E, Donnelly S, Christiansen MH. Individual differences in language acquisition and
364	processing. Trends in cognitive sciences. 2018;22(2):154-69. doi:
365	10.1016/j.tics.2017.11.006
366	2. Hamrick P. Declarative and procedural memory abilities as individual differences in
367	incidental language learning. Learning and Individual Differences. 2015;44:9-15. doi:
368	10.1016/j.lindif.2015.10.003.
369	3. Ruiz S, Tagarelli KM, Rebuschat P. Simultaneous acquisition of words and syntax: Effects
370	of exposure condition and declarative memory. Frontiers in Psychology. 2018 12;9:1168.
371	doi: 10.3389/fpsyg.2018.01168

- 4. Li S. Cognitive differences and ISLA. In: Loewen S, Sato M, editors. The Routledge
 handbook of instructed second language acquisition. New York: Routledge; 2017. pp. 396417.
- 5. Pawlak M. Overview of learner individual differences and their mediating effects on the
 process and outcome of interaction. In Gurzynski-Weiss L., editor. Expanding individual
 difference research in the interaction approach: Investigating learners, instructors, and
 other interlocutors. Amsterdam: John Benjamins; 2017. pp. 19-40.
- 379
 6. Larsen- Freeman D. Looking ahead: Future directions in, and future research into, second
 380 language acquisition. Foreign language annals. 2018;51(1):55-72. doi: 10.1111/flan.12314
- 7. Hamrick P, Lum JA, Ullman MT. Child first language and adult second language are both
 tied to general-purpose learning systems. Proceedings of the National Academy of
 Sciences. 2018;115(7):1487-92.
- 384
 8. Lado B. Aptitude and pedagogical conditions in the early development of a nonprimary
 385
 language. Applied Psycholinguistics. 2017;38(3):679-701.
- Faretta-Stutenberg M, Morgan-Short K. The interplay of individual differences and context
 of learning in behavioral and neurocognitive second language development. Second
 Language Research. 2018;34 (1): 67-101. doi: 10.1177/0267658316684903.
- 10. Tagarelli KM, Ruiz S, Moreno Vega JL, Rebuschat P. Variability in second language
 learning: The roles of individual differences, learning conditions, and linguistic
 complexity. Studies in Second Language Acquisition. 2016;38(2):293-316. doi:
 10.1017/S0272263116000036.
- 393 11. Buffington J, Morgan-Short K. Declarative and procedural memory as individual
 394 differences in second language aptitude. In: Wen Z, Skehan P, Biedroń A, Li S, Sparks R,

- 395 editors. Language aptitude: Multiple perspectives and emerging trends. New York:
 396 Routledge; 2019. pp. 215–237.
- Marsden E, Morgan- Short K, Thompson S, Abugaber D. Replication in second language
 research: Narrative and systematic reviews and recommendations for the field. Language
 Learning. 2018;68(2): 321-91. doi:10.1111/lang.12286.
- 400 13. Plonsky L. Study quality in SLA: An assessment of designs, analyses, and reporting
 401 practices in quantitative L2 research. Studies in Second Language Acquisition. 2013;35(4):
 402 655-87. doi: 10.1017/S0272263113000399
- 403 14. Plonsky L. Quantitative research methods. In: Loewen S, Sato M, editors. The Routledge
 404 handbook of instructed second language acquisition. New York: Routledge; 2017. pp. 505405 521.
- 406 15. Lindstromberg S. Inferential statistics in Language Teaching Research: A review and ways
 407 forward. Language Teaching Research. 2016;20(6): 741-68. doi:
 408 10.1177/1362168816649979.
- 409 16. Krantz JH, Reips UD. The state of web-based research: A survey and call for inclusion in
 410 curricula. Behavior Research Methods. 2017;49(5): 1621-1619. doi: 10.3758/s13428-017411 0882-x
- 412 17. Hicks KL, Foster JL, Engle RW. Measuring working memory capacity on the web with
 413 the online working memory lab (the OWL). Journal of Applied Research in Memory and
 414 Cognition. 2016;5(4): 478-89. doi: 10.1016/j.jarmac.2016.07.010.
- 415 18. Wolfe CR. Twenty years of Internet-based research at SCiP: A discussion of surviving
 416 concepts and new methodologies. Behavior research methods. 2017;49(5): 1615-1620. doi:
- 417 10.3758/s13428-017-0858-x.

418	19. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil
419	administration of patient-reported outcome measures: a meta-analytic review. Value in
420	Health. 2008;11(2): 322–333. doi: 10.1111/j.1524-4733.2007.00231.x.
421	20. Cowan N. Working memory maturation: Can we get at the essence of cognitive growth?.
422	Perspectives on Psychological Science. 2016;11(2): 239-64. doi:
423	10.1177/1745691615621279
424	21. Baddeley AD. Modularity, working memory and language acquisition. Second Language
425	Research. 2017;33(3): 299-311. doi: 10.1177/0267658317709852.
426	22. Roehr K. Linguistic and metalinguistic categories in second language learning. Cognitive
427	Linguistics. 2008;19(1): 67-106. doi: 10.1515/COG.2008.005.
428	23. Grundy JG, Timmer K. Bilingualism and working memory capacity: A comprehensive
429	meta-analysis. Second Language Research. 2017;33(3): 325-40. doi:
430	10.1177/0267658316678286.
431	24. Jeon EH, Yamashita J. L2 reading comprehension and its correlates: A meta- analysis.
432	Language Learning. 2014;64(1):160-212. doi: 10.1111/lang.12034.
433	25. Linck JA, Osthus P, Koeth JT, Bunting MF. Working memory and second language
434	comprehension and production: A meta-analysis. Psychonomic Bulletin & Review.
435	2014;21(4): 861-83. doi: 10.3758/s13423-013-0565-2.
436	26. Bailey H, Dunlosky J, Kane MJ. Contribution of strategy use to performance on complex
437	and simple span tasks. Memory & cognition. 2011;39(3): 447-61. doi: 10.3758/s13421-
438	010-0034-3.
439	27. Turner ML, Engle RW. Is working memory capacity task dependent?. Journal of memory
440	and language. 1989;28(2): 127-54. doi: 10.1016/0749-596X(89)90040-5.

- 28. Conway ARA, Kane, MJ, Bunting MF, Hambrick DZ, Wilhelm O, et al. (2005) Working
 memory span tasks: A methodological review and user's guide. Psychonomic Bulletin and
 Review 12(12): 769–786. doi: 10.3758/BF03196772.
- 444 29. Zhou H, Rossi S, Chen B. Effects of working memory capacity and tasks in processing L2
- 445 complex sentence: evidence from Chinese-English bilinguals. Frontiers in psychology.
 446 2017;8: 595. doi: 10.3389/fpsyg.2017.00595
- 30. Reber PJ, Knowlton BJ, Squire LR. Dissociable properties of memory systems: differences
 in the flexibility of declarative and nondeclarative knowledge. Behavioral Neuroscience.
 1996;110(5): 861. doi: 10.1037/0735-7044.110.5.861.
- 450 31. Squire LR. Memory systems of the brain: a brief history and current perspective.
 451 Neurobiology of learning and memory. 2004;82(3): 171-7. doi: 10.1016/j.nlm.2004.06.005
- 452 32. Eichenbaum H. Hippocampus: cognitive processes and neural representations that underlie
 453 declarative memory. Neuron. 2004;44(1):109-20.
- 454 33. Squire LR. Memory systems of the brain: a brief history and current perspective.
 455 Neurobiology of learning and memory. 2004;82(3): 171-7. doi:
 456 10.1016/j.nlm.2004.06.005.
- 457 34. Knowlton BJ, Siegel AL, Moody TD. Procedural learning in humans. In Byrne JH, editor.
 458 Learning and memory: A comprehensive reference. 2nd ed. Oxford: Academic Press;
 459 2017. pp. 295–312.
- 35. Hamrick P, Lum JA, Ullman MT. Child first language and adult second language are both
 tied to general-purpose learning systems. Proceedings of the National Academy of
 Sciences. 2018;115(7): 1487-1492. doi: 10.1073/pnas.1713975115.

- 36. Ullman MT. The declarative/procedural model: A neurobiologically motivated theory of
 first and second language. In: VanPatten B, Williams J, editors. Theories in second
 language acquisition: An introduction. 2nd ed. New York: Routledge; 2015. pp. 135-158.
- 466 37. Ullman MT. The declarative/procedural model: A neurobiological model of language
 467 learning, knowledge, and use. In: Hickok G, Small SA, editors. Neurobiology of language.
 468 Amsterdam: Elsevier; 2016. pp. 498–505.
- 38. Morgan-Short K, Faretta-Stutenberg M, Brill-Schuetz KA, Carpenter H, Wong PC.
 Declarative and procedural memory as individual differences in second language
 acquisition. Bilingualism: Language and Cognition. 2014;17(1):56-72. doi:
 10.1017/S1366728912000715.
- 473 39. Carpenter, HS. A behavioral and electrophysiological investigation of different aptitudes
 474 for L2 grammar in learners equated for proficiency level. Ph.D. Thesis, Georgetown
 475 University. 2008. Available from: http://hdl.handle.net/10822/558127.
- 476 40. Carroll JB, Sapon SM. Modern Language Aptitude Test: Manual. New York:
 477 Psychological Corporation; 1959.
- 478 41. Trahan DE, Larrabee GJ. Continuous visual memory test. Odessa, FL: Assessment
 479 Resources. 1988.
- 480 42. Schneider W, Eschman A, Zuccolotto A. EPrime user's guide. Pittsburgh, PA: Psychology
 481 Software Tools Inc. 2002.
- 43. Unsworth N, Heitz RP, Schrock JC, Engle RW. An automated version of the operation
 span task. Behavior Research Methods. 2005;37(3): 498-505. doi: 10.3758/BF03192720.
- 484 44. Wickens TD. Elementary signal detection theory. New York: Oxford University Press;
 485 2002.

- 486 45. Kane MJ, Hambrick DZ, Tuholski SW, Wilhelm O, Payne TW, Engle RW. The generality
 487 of working memory capacity: a latent-variable approach to verbal and visuospatial memory
 488 span and reasoning. Journal of Experimental Psychology: General. 2004;133(2): 189-217.
 489 doi: 10.1037/0096-3445.133.2.189.
- 46. Marsden E, Morgan- Short K, Thompson S, Abugaber D. Replication in second language
 research: Narrative and systematic reviews and recommendations for the field. Language
 Learning. 2018;68(2): 321-391. doi: 10.1111/lang.12286.
- 493 47. Marsden E, Morgan- Short K, Thompson S, Abugaber D. Replication in second language
 494 research: Narrative and systematic reviews and recommendations for the field. Language
 495 Learning. 1;68(2): 321-391. doi: 10.1111/lang.12286.
- 496 48. Gelman A. The failure of null hypothesis significance testing when studying incremental
 497 changes, and what to do about it. Personality and Social Psychology Bulletin. 2018;44(1):
 498 16-23. doi: 10.1177/0146167217729162.
- 499 49. Leidheiser W, Branyon J, Baldwin N, Pak R, McLaughlin A. Lessons learned in adapting
- 500a lab-based measure of working memory capacity for the web. In: Proceedings of the501Human Factors and Ergonomics Society Annual Meeting 2015. Los Angeles: Sage CA;
- 502 2015. pp. 756-760.
- 503 50. MacWhinney B. A shared platform for studying second language acquisition. Language
 504 Learning. 2017;67(S1): 254-75. doi: 10.1111/lang.12220.
- 505 51. Meurers D, Dickinson M. Evidence and interpretation in language learning research:
 506 Opportunities for collaboration with computational linguistics. Language Learning.
 507 2017;67(S1): 66-95. doi: 10.1111/lang.12233.

508	52. Ziegler N, Meurers D, Rebuschat P, Ruiz S, Moreno- Vega JL, Chinkina M, Li W, Grey
509	S. Interdisciplinary research at the intersection of CALL, NLP, and SLA: Methodological
510	implications from an input enhancement project. Language Learning. 2017;67(S1): 209-
511	231. doi: 10.1111/lang.12227.

94 95	Measuring individual differences in cognitive abilities in the lab and on the web
95 96	
97	Simón Ruiz ^{1*} , Xiaobin Chen ¹ , Patrick Rebuschat ^{1,2} Detmar Meurers ^{1,3}
98	
99	
100 101	
101	¹ [¶] LEAD Graduate School and Research Network, University of Tübingen, Tübingen, Germany
103	
104	² [¶] Department of Linguistics and English Language, Lancaster University, Lancaster, United
105	Kingdom
106	
107	³ Department of Linguistics, University of Tübingen, Tübingen, Germany
108	
109	* Corresponding author
110	E-mail: simon.ruiz-hernandez@sfs.uni-tuebingen.de (SR)
111	
112	
113	[¶] These authors contributed equally to this work.
114	
115	

116 Abstract

117 The present study compared lab-based and web-based versions of cognitive individual 118 difference measures widely used in second language research (working memory and declarative 119 memory). Our objective was to validate web-based versions of these tests for future research and 120 to make these measures available for the wider second language research community, thus 121 contributing to the study of individual differences in language learning. The establishment of 122 measurement equivalence of the two administration modes is important because web-based testing 123 allows researchers to address methodological challenges such as restricted population sampling, 124 low statistical power, and small sample sizes. Our results indicate that the lab-based and web-125 based versions of the tests were equivalent, i.e., scores of the two test modes correlated. The 126 strength of the relationships, however, varied as a function of the kind of measure, with 127 equivalence appearing to be stronger in both the working memory and the verbal declarative 128 memory tests, and less so in the nonverbal declarative memory test. Overall, the study provides 129 evidence that web-based testing of cognitive abilities can produce similar performance scores as 130 in the lab.

131

132 Introduction

Individual differences can greatly affect how we acquire and process language [1-3] and mediate and/moderate the effectiveness of instruction [4]. In adult language learning, for example, learners' cognitive abilities have great explanatory power in accounting for differences in learning outcomes ([5-6]). For instance, working memory and declarative memory are considered to be particularly important sources of learner variation (e.g., [7-10]; see [4, 11], for reviews).

138 The effect of working memory and declarative memory on language learning has been 139 primarily studied in lab settings, i.e., in well-controlled environments where participants are tested 140 individually. While this choice is methodologically sound, it can also negatively affect sample size 141 and population sampling [12, 13, 14]. Lab-based testing generally means testing participants 142 individually and sequentially, which is labor-intensive and could explain why lab studies tend to 143 have (too) few participants to allow for meaningful generalization. As an example, in second 144 language (L2) research, Plonsky [13] found that the typical sample size in L2 studies was 19 145 participants, and Lindstromberg [15] recently reported a similar small average sample size of 20 146 participants. In the same vein, [16] reported that, in psychology, median sample sizes have not 147 increased considerably in the last two decades, and are generally too small to detect small effect 148 sizes, which are distinctive of many psychological effects. Moreover, many (if not most) lab 149 studies in research draw their sample from the surrounding student population, which is 150 understandable given the ease of access, but also means that samples are often not representative 151 of the population of interest. Conducting research by means of remote testing via the web could 152 alleviate some of these concerns. For example, web-based testing facilitates the acquisition of large 153 amounts of data since participants can be tested simultaneously, enabling researchers to run higher-154 powered studies. Likewise, test administration can also be more cost-effective than research 155 conducted in the lab [17].

The use of (remote) web-based testing can also offer other important methodological advantages over other forms of simultaneous delivery of tests, such as traditional paper-pencil and (offline) computer-based testing [18, 19]. Particularly, it allows researchers to standardize and optimize testing procedures, which can contribute to more consistent and uniform test-taking conditions across different locations and times [20]. This, in turn, can also facilitate the replication of studies [21]. Moreover, remote testing via the web can reduce experimenter effects, as testing can occur in more ecologically-valid settings, and without any direct contact between experimenters and participants [20, 21]. Finally, and more importantly, web-based experimenting has been found to be a reliable and effective research tool [17, 22, 23].

165 The present study compared lab-based and web-based versions of cognitive tests that are 166 widely used in disciplines such as psychology and second language research. Particularly, our 167 intent was to compare performance of measures as they are originally used in the lab with their 168 corresponding online versions. In doing so, our objective was to validate the web-based tests for 169 use in subsequent research and to make these available to the wider research community, and 170 especially to researchers working on the area of L2 acquisition. The sharing of tasks, especially of 171 tasks that permit the collection of substantial amounts of data via the web, will be an important 172 component in alleviating the data collection issues associated with lab-based research. Moreover, 173 making these specific tasks available will also contribute directly to our understanding of 174 individual differences in L2 acquisition. To support such task sharing and use, it is essential to first 175 establish the validity of the online versions of the tasks (on a par with what is established about 176 the offline versions). With this in mind, the study set out to establish measurement equivalence 177 between lab-based and web-based tests of working memory and declarative memory.

According to Gwaltney, Shields and Shiffman ([24], p. 323), measurement equivalence can be established if "1) the rank orders of scores of individuals tested in alternative modes closely approximate each other; and 2) the means, dispersions, and shapes of the score distributions are approximately the same". The first type of equivalence regards to whether differences observed in one measurement are also systematically found in the other, meaning that, even when the two measurements produce two different numbers, these numbers are clearly and systematically 184 associated with each other. The second type concerns whether two measurements yield the same 185 numbers. Considering that this study is a subcomponent of the dissertation research of the first 186 author, limiting funding and time (see limitations below), we focused the investigation on one type 187 of measurement equivalence, the first type: Do people who have relatively high values in one of 188 tests also have relatively high values on the other test, and the other way around? More specifically, 189 we compare the differential performance generated by two versions of tests measuring working 190 memory and declarative memory abilities in lab-based and web-based settings, in order to assess 191 whether the two versions are equivalent regarding the relationships between scores.

192 Assessing equivalence between lab and web-based measurements is essential for several 193 reasons. Firstly, it is necessary to demonstrate that the findings obtained in web-based studies are 194 comparable to those of previous research, which have been mainly collected in lab-based settings. 195 Secondly, it is important to ensure that cognitive constructs are similarly gauged in both testing 196 modalities. Likewise, it is crucial to establish whether lab-based and web-based tests are 197 equivalent, given that web-based testing could prove to be a viable way to tackle some of the 198 current methodological issues found in research conducted in lab-based settings, such as 199 underpowered studies, restricted population sampling, and small sample sizes [17, 22, 23]. Of 200 these methodological issues, in particular, low statistical power and small sample sizes have been 201 identified as key factors in the ongoing discussions about the reproducibility of research findings 202 in life and social sciences [25-27]. In psychology, for example, there is currently considerable 203 debate about the so-called *replication crisis* [28], that is, failure to reproduce significant findings 204 when replicating previous research [27]. In this regard, and considering that much research is 205 underpowered [29, 30], web-based testing can enable the collection of larger sample sizes, and 206 thus contribute to achieve more statistical power to detect the effects of interest. On the other hand,

the ease of access, cost-effectiveness, and practicality of web-testing can also increase the attempts
to reproduce results from previous studies, and thus making (large-scale) replication studies more
appealing for researchers to undertake [30].

210

211 Working memory

Working memory is the capacity to process and hold information at the same time while performing complex cognitive tasks such as language learning, comprehension and production [31]. According to Baddeley and colleagues (e.g., [32]), working memory is a multicomponent system that includes storage subsystems responsible for retaining both visual-spatial and auditory information, an episodic buffer that serves as a link between the storage subsystems and long-term memory, and a central executive that acts as an attentional control system.

218 Regarding L2 learning, working memory assists learners to simultaneously process form, 219 meaning and use of language forms. More specifically, working memory is involved in key 220 cognitive processes such as decision making, attention control, explicit deduction, information 221 retrieval and analogical reasoning [4]. Moreover, working memory is also important for retaining 222 metalinguistic information while comprehending and producing L2 language [33]. In this regard, 223 meta-analytic work has reported the important role of working memory in L2 comprehension and 224 production (e.g., [34-36]). For example, Linck et al. ([36], p. 873) found that working memory has 225 a positive impact on L2 comprehension outcomes (r = 0.24). Likewise, Jeon and Yamashita's [35] 226 meta-analysis also showed that working memory is related to L2 reading comprehension (r = 0.42). 227 Regarding production, meta-analytic research has, too, indicated a significant association with 228 working memory (e.g., [36]). In this case, Linck et al. ([36], p. 873) found a positive correlation 229 for productive outcomes as well (r = 0.27).

7

230 Working memory is often measured by means of simple or complex span tasks. Simple 231 span tasks, such as digit span and letter span, entails recalling short lists of items, and they seek to 232 measure the storage component of working memory [37]. Complex span tasks, such as the 233 operation span task (OSpan; [38]), on the other hand, include remembering stimuli while 234 performing a another task. This type of tasks taxes both processing (attention) and storage 235 (memory) aspects of working memory [32]. Here, we focus on a complex task, namely the OSpan. 236 This complex task has been found to be a valid and reliable measure of working memory capacity 237 [39], and has also been recommended as a more accurate measure to examine the association 238 between working memory and L2 processing and learning [40].

239

240 **Declarative memory**

Declarative memory is the capacity to consciously recall and use information [41]. The declarative memory system is one of the long-term memory systems in the brain [42]. It is mainly responsible for the processing, storage, and retrieval of information about facts (semantic knowledge) and events (episodic knowledge; [43, 44]). Learning in the declarative memory system is quick, intentional, and attention-driven [45].

Substantial research has now investigated the role of declarative memory in first and second language acquisition [46]. In first language acquisition, declarative memory is involved in the processing, storage and learning of both arbitrary linguistic knowledge (e.g., word meanings) as well as rule-governed aspects of language (e.g., generalizing grammar rules [47, 48]). In the case of L2 acquisition, declarative memory underpins the learning, storage and processing of L2 vocabulary and grammar [47, 48], at least in the earliest phases of acquisition [46, 49]. Several studies (e.g., [2, 9, 49, 50]) has confirmed the predictive ability of declarative memory to explain
variation in L2 attainment.

Declarative memory has been tested through recall and recognition tasks (e.g., 49, 50), both verbal, such as the paired associates subtest of the Modern Language Aptitude Test (MLAT5; [51]), and nonverbal, such as the Continuous Visual Memory Task (CVMT; [52]).

257

258 **The present study**

259 The main goal of the present study was to provide web-based versions of commonly employed 260 individual difference measures in second language research, in order to make them usable in large-261 scale intervention studies (generally in authentic, real-life learning contexts). To that end, we 262 examined whether lab-based and web-based versions of working memory and declarative memory 263 tests yield similar performance scores, i.e., whether the two versions were equivalent or 264 comparable. More specifically, we assessed whether the values of one type of mode of 265 administration corresponded to the values in the other mode (i.e., first type of equivalence). In 266 other words, are the differences in scores constant, or parallel in the two ways of measuring? The 267 web-based versions are freely available; to use the test, please send an email to the first author.

268

269 Methods

270 Ethics statement

This research was approved by the Commission for Ethics in Psychological Research, University of Tübingen, and all participants provided written informed consent prior to commencement of the study.

275 **Participants**

276 Fifty participants (37 women and 13 men), with a mean age of 26.4 years (SD = 4.2), 277 partook in the study. The majority of participants were native speakers of German (72%), followed 278 by Russian (8%), Spanish (6%), Chinese (4%), English, Hungarian, Persian, Serbian and 279 Vietnamese (2% each). Seven (14%) participants did not complete the second half of the study 280 (i.e., web-based testing). Additionally, participant numbers differed across test versions due to 281 technical difficulties (i.e., participants entered their responses using the wrong keys [Web-based 282 CVMT]; and data was not correctly saved for one participant [Web-based MLAT5]; see 283 description and Table 1 below, and Discussion). Twenty-seven participants were graduate students 284 (54%), and twenty-three were undergraduates (46%). Participants self-reported English 285 proficiency, with most being advanced learners (82%), followed by intermediate (18%). All 286 subjects gave informed consent and received €20 for participating.

287 Materials

288 Three cognitive tests were administered, one measuring working memory capacity, and 289 two assessing verbal and nonverbal declarative memory abilities, respectively. In the lab-based 290 setting, both working memory and nonverbal declarative memory tests were programmed and 291 delivered via E-Prime v2.0 [53]; the verbal declarative memory test was given in paper-pencil 292 form, as originally developed and delivered. Moreover, web-based versions of the three 293 cognitive tests were developed for this study using Java with the GoogleWeb Toolkit 294 (http://www.gwtproject.org), and were accessible from all browsers. A description of each test is 295 given below.

297 Working memory. An adapted version of the Automated Operation Span Task (OSpan; [54]), a 298 computerized form of the complex span task created by Turner and Engle [38], was used to 299 gauge participants' working memory capacity [9, 22]. Based on the Klingon Span Task 300 implemented by Hicks et al. [22], this version consisted of using Klingon symbols instead of 301 letters, the stimuli to be remembered in the original OSpan task. In Hicks et al.' study, 302 participants cheated by writing down the letter memoranda in the web-based version of the 303 classic OSpan, motivating the change of the original stimuli. The task included a practice phase 304 and a testing phase. In the practice phase, participants were first shown with a series of Klingon 305 symbols on the screen, and then were asked to recall them in the order in which they had 306 appeared after each trial (i.e., symbol recall). Next, participants were required to solve a series of 307 simple equations (e.g., 8 * 4 + 7 = ?). Finally, subjects performed the symbol recall while also 308 solving the math problems, as they would later do in the actual testing phase. Following the 309 practice phase, participants were shown with the real trials, which consisted of a list of 15 sets of 310 3–7 randomized symbols that appeared intermingled with the equations. In sum, there were 75 311 symbols and 75 math problems. At the end of each set, participants were asked to remember the 312 symbols in the sequence they had been presented. An individual time limit to answer the math 313 problems in the real trials was calculated from the average response time plus 2.5 standard 314 deviations taken during the math practice section. Following Unsworth et al. [54], a partial score 315 (i.e., total number of correct symbols recalled in the correct order) was taken as the OSpan score 316 (see [39], for a description of scoring procedures). The highest possible score was 75. The entire 317 task took about 25 min.

319 **Verbal declarative memory.** To measure verbal declarative memory, the Modern 320 Language Aptitude Test, Part 5, Paired Associates (MLAT5; [51]), was used [9, 49, 50]. In the 321 MLAT5, participants were required to memorize artificial, pseudo-Kurdish words and their 322 meanings in English. Participants were first asked to study 24-word association pairs for two 323 minutes, and then complete a two-minute practice section. The list of foreign words with their 324 respective English meanings was made available for participants as they completed the practice 325 session. Finally, subjects were instructed to complete a timed multiple-choice test (four minutes), 326 by selecting the English meaning of each of the 24 pseudo-Kurdish words from five options 327 previously displayed at the memorization stage. For each correct response, one point was given, 328 yielding a total score of 24 points. The test duration was about 8 minutes.

329

Nonverbal declarative memory. The Continuous Visual Memory Task (CVMT; [52]) 330 331 served as a measure of nonverbal declarative memory [9, 49, 50]. As a visual recognition test, 332 the CVMT is entails asking participants to first view a collection of complex abstract designs on 333 the screen, and then to indicate whether the image they just saw was novel ("new") in the 334 collection, or they had seen the image before ("old"). Seven of the designs were "old" (target 335 items), and 63 were "new" (distractors). The target items appeared seven times (49 trials), and 336 the distractors only once (63 trials) across the test. All items were shown in a random but fixed 337 order, each one appearing on the screen for two seconds. Following the two seconds, participants 338 were instructed to respond to the "OLD or NEW?" prompt on the screen. In the lab-based mode, 339 subjects used mouse click for making their choice, left for "NEW", or right for "OLD". In the web-based mode, they responded by pressing either the "N" key for "NEW", or the "O" key for 340

341 "OLD" on the keyboard. The CVMT took 10 min to complete. A *d*'(d-prime) score [55] was
342 calculated for each participant. The d' score was used to reduce potential response bias.

343

344 **Procedure**

As previously noted, participants underwent two cognitive testing sessions, one in the lab and one on the web. In the lab-based session, with the assistance of a proctor, each subject was tested individually. After providing informed consent, participants took the three cognitive tests under investigation in fixed order: OSpan, CVMT, and MLAT5. Upon finishing the MLAT5, subjects then filled out a background questionnaire. The whole lab-based session lasted about 40 min.

351 Regarding the web-based session, each subject was sent an email with a unique web link 352 with a personalized code, which once clicked, took them to an interface that hosted the web-based 353 versions of the cognitive tests. In order to avoid multiple responses by the same participant, the 354 link was disabled once subjects had submitted their responses in the last test (i.e., MLAT5). In the 355 email, participants were also informed that the web-based session lasted about 40 min, and that it 356 had to be completed within a week. On the interface, following informed consent, subjects were 357 provided with general instructions that reflected the nature of a web-based experiment. Such 358 instructions included completing the experiment in a quiet place without interruption, and from 359 start to finish in one sitting. Likewise, the use of the browser's back button, refreshing the browser 360 page, or closing the browser window were prohibited. Importantly, participants were instructed 361 not to take any notes at any point during the entire experiment. The web-based tests were given in 362 the same fixed order as in the lab-based session. On average, the mean period between the first 363 and second testing was 45.7 days (SD = 4.1).

364 **Results**

365 All data were analyzed by means of R (version 3.3.2; [56]). Missing data was ignored 366 (complete-case analysis). Linear regression models were built using the lm function in the lme4 367 library [57]. From a temporal perspective, lab scores were used to predict web scores in the 368 linear regression models. To verify normality, model residuals were visually inspected. 369 Reliability was assessed using Cronbach's alpha. Following Kane et al. [58], for the lab-based 370 working memory test (OSpan-Lab-based), reliability was assessed by calculating the proportion 371 of correctly recalled Klingon symbols per each of the 15 trials in the test (e.g., one out of four 372 symbols correctly recalled corresponded to a proportion of .25). For the web-based working 373 memory test (OSpan-Web-based), however, internal consistency is not reported, since it was not 374 technically possible to perform a detailed item-based analysis. Descriptive statistics are presented 375 first, followed by correlations, internal consistency estimates (Cronbach's alpha), and the results 376 of linear regression analyses.

377

Descriptive statistics

379 Table 1 presents the descriptive statistics for participants' performance on cognitive tests380 in both testing settings.

Test	Ν	М	SD	Skew	Kurtosis
OSpan Lab-based	50	25.78	13.34	0.61	2.90
OSpan Web-based	43	29.79	15.42	0.67	3.26
MLAT5 Lab-based	50	17.92	5.50	-0.64	2.49
MLAT5 Web-based	42	19.10	5.81	-1.19	3.58
CVMT Lab-based	49	1.99	0.46	0.23	3.35
CVMT Web-based	40	2.30	0.63	0.73	3.32

382 Table 1. Descriptive statistics for comparison of lab-based and web-based testing.

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task.

383

384 Correlations

385 Table 2 and Fig 1 show the correlations between/among the different versions of the

individual difference tests.

387

389 Table 2. Correlations between lab-based and web-based scores for individual difference

390 **tests.**

Test	OSpan	OSpan	MLAT5	MLAT5	CVMT
	Lab-based	Web-based	Lab-based	Web-based	Lab-based
OSpan Web-based	.80				
MLAT5 Lab-based	.40	.51			
MLAT5 Web-based	.32	.40	.82		
CVMT Lab-based	.19	.31	.42	.23	
CVMT Web-based	.21	.30	.21	.19	.55

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task.

391

392 Fig 1. Scatterplots of the correlation of each pair of lab-based and web-based versions of

- 393 individual difference measures.
- 394 OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern
- 395 Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous
- 396 Visual Memory Task.
- 397

398 Reliability

- 399 Table 3 presents Cronbach's alpha values of individual test versions.
- 400
- 401

Test	Cronbach's alpha
OSpan Lab-based	.86
MLAT5 Lab-based	.77
MLAT5 Web-based	.93
CVMT Lab-based	.63
CVMT Web-based	.67

402 **Table 3. Cronbach's alphas for cognitive test versions.**

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task.

403

404 **Regression analysis**

405 The results of the regression analyses are displayed in Table 4. For the working memory

406 test (OSpan), the unstandardized coefficient was .89 ($\beta = .77$, SE = 0.10, p < .001). For the verbal

407 declarative memory test (MLAT5), the unstandardized coefficient was .83 (β = .78, SE = 0.09, p

408 <.001). And for the nonverbal declarative memory test (CVMT), the unstandardized coefficient

409 was .74 ($\beta = .54$, SE = 0.19, p < .001). Overall, the results indicated that the lab-based and web-

410 based scores are substantially related.

411

Test	Unstandardized coefficient ^a	SE	р
OSpan	0.89 (.77)	0.10	<.001
MLAT5	0.83 (.78)	0.09	< .001
CVMT	0.74 (.54)	0.19	< .001

413 **Table 4. Regression for comparison of lab-based and web-based scores.**

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task. ^aThe standardized coefficient (β) in parentheses.

414

415 **Discussion**

416 Studies on individual differences in language learning frequently assess the working 417 memory and declarative memory capacities of their participants in order to determine the effect 418 of these cognitive variables on learning outcomes. Most of this research, however, is conducted 419 in lab-based settings, which often implies relatively small sample size and a restricted population 420 sampling. Both of these methodological challenges can be addressed by means of remote testing 421 via the web. In the present study, we compared lab-based and web-based individual difference 422 measures in order to validate web-based tests for future research. The type of comparison 423 contributes significantly to ongoing efforts to improve the methodological robustness of current 424 second language research, for example [12]. If web-based testing can be shown to yield 425 comparable results to lab-based testing, researchers will be able to reach more participants for 426 their studies, which, in turn, can help alleviate some of the current concerns in lab-based research 427 (e.g., low statistical power, non-representative population samples, and small sample sizes). In

addition, demonstrating the equivalence of lab-based and web-based measures of the same
individual difference constructs is essential for the comparability of results across studies.
Crucially, establishing measurement equivalence between lab-based and web-based versions will
also provide assurance that the tests are measuring cognitive constructs the same way regardless
of administration mode [17, 59].

433 Findings showed that the scores in the lab-based and web-based versions of three 434 cognitive tests (MLAT5, CVMT, OSpan) were equivalent concerning differences in 435 performance, which were constant in the two versions, suggesting that participants who had 436 relatively high values in one task also had relatively high values in the second, or the other way 437 around. However, the strength of the relationship was a function of the kind of test. More 438 specifically, in both the working memory test (OSpan) and the verbal declarative memory test 439 (MLAT5), the scores were more strongly correlated ($\beta = .77$ and $\beta = .78$, respectively); for the 440 nonverbal declarative test (CVMT), equivalence appears to be weaker ($\beta = .54$). Overall, the 441 correlations reported here between lab-based and web-based scores are consistent with the 442 assumption that both versions seem to likely measure the same cognitive construct, at least for 443 the working memory test (OSpan) and the verbal declarative memory test (MLAT5), and, to a 444 lesser extent, for the nonverbal declarative test (CVMT).

A potential explanation for lesser equivalence in the versions of the nonverbal declarative test (CVMT) could be due to the different manner in which the responses to the visual stimuli were entered in the two testing modes. It will be recalled that in the lab-based version participants used left ("NEW") or right ("OLD") mouse clicking to provide a response, whereas in the web-based version, they used the keyboard ("N" and "O" keys). This modification made to

450 the web-based version was motivated by technical reasons, specifically, the browser window 451 may not register the participants' response if the cursor is not over a certain area on the page, 452 which in turn may cause problems of missing data. Previous research has found that participants 453 in web-based research are particularly prone to err when using the keyboard to input their 454 responses [60], which in this case might have affected the results of the comparison between lab-455 based and web-based versions of CVMT. Future research comparing performance between the 456 lab and web-versions may benefit from collecting data through touch input instead, as this might 457 help overcome potential technical difficulties caused by using mouse clicking for web-based 458 data.

459 Some limitations of the study and the findings presented here should be considered. One 460 of the limitations was the small sample size. As mentioned earlier, logistic constrains due to the 461 availability of time and funding prevented the researchers from testing more participants for this 462 study. In addition, the fact that some participants (14%) dropped out before completing any of 463 the web-based measures in the second part of the experiment, which is typical in web-based 464 research [17], also contributed to the reduction of the data available for the comparison between 465 lab and web-based testing in the present investigation. Therefore, our findings should be 466 replicated in a larger study. A second limitation was that test-retest reliability was not examined 467 here, given that the main aim of this study was to establish valid online versions of known 468 individual difference measures. Future research should assess test-retest reliability, as it is as an 469 interesting endeavor for studying individual difference measures in future work. Finally, and as 470 indicated above, a third limitation concerned technical issues that affected data collection, as 471 some participants used the wrong keys on the keyboard to submit their responses to the web-472 based version of the CVMT, rendering the data from some of the participants impossible to use

473 for the comparison; furthermore, data from one subject was missing in the Web-based MLAT, 474 which may have been due to technical issues at the participant's end (e.g., not following the 475 general instructions given, such as refreshing or closing the browser page [see Procedure]; or 476 Internet disconnection). In this sense, Reips and Krantz [61] (see also[17]) caution researchers 477 that one of the potential disadvantages of Internet-driven testing is the technical variability 478 characteristic of web-based research (e.g., different browsers and Internet connections), which, in 479 turn, may affect data collection.

480 Conclusion

481 This study aimed to establish the validity of using web-based versions of established 482 offline tasks. As such, the study has provided evidence that it is possible to measure individual 483 differences in cognitive abilities on the web and obtain similar performance as in the lab. The 484 lab-based and web-based versions of the three cognitive tests are comparable or equivalent. 485 However, given that they do not perfectly correlate, we recommend using one of the two modes 486 within one study and not comparing individual scores from one mode with scores from the other. 487 Moreover, the extent to which the measures are equivalent varies according to the test. In this 488 sense, we are confident that the two versions for the working memory test (OSpan) and the 489 verbal declarative memory (MLAT5) are likely to measure the same construct, whereas the 490 correlation between the nonverbal declarative test (CVMT) versions was less pronounced. Our 491 research has shown that collecting experimentally controlled data on cognitive individual 492 differences typically used in the area of L2 research in the Internet is feasible and comparable to 493 lab-based collection. Consequently, some of these web-based versions could very well be

494	incorporated, for example, in future web-based intervention studies on second language learning,
495	thereby contributing to the scaling up of data collection in the field [62-64].

496

497 Acknowledgements

498

We would like to thank Johann Jacoby, for his invaluable suggestions that strengthened ourexperimental design and analysis.

501

502 **References**

503	1.	Kidd E, Donr	nelly S, Chr	istians	en MH. Indivi	dual differenc	es in language acquisiti	on and
504		processing.	Trends	in	Cognitive	Sciences.	2018;22(2):154-69.	doi:
505		10.1016/j.tics	.2017.11.00)6				

- Hamrick P. Declarative and procedural memory abilities as individual differences in incidental language learning. Learning and Individual Differences. 2015;44:9-15. doi: 10.1016/j.lindif.2015.10.003.
- 3. Ruiz S, Tagarelli KM, Rebuschat P. Simultaneous acquisition of words and syntax: Effects
 of exposure condition and declarative memory. Frontiers in Psychology. 2018 12;9:1168.
- 511 doi: 10.3389/fpsyg.2018.01168
- 4. Li S. Cognitive differences and ISLA. In: Loewen S, Sato M, editors. The Routledge
 handbook of instructed second language acquisition. New York: Routledge; 2017. pp. 396417.
- 515 5. Pawlak M. Overview of learner individual differences and their mediating effects on the
 516 process and outcome of interaction. In Gurzynski-Weiss L., editor. Expanding individual

- 517 difference research in the interaction approach: Investigating learners, instructors, and 518 other interlocutors. Amsterdam: John Benjamins; 2017. pp. 19-40. 519 6. Larsen- Freeman D. Looking ahead: Future directions in, and future research into, second 520 language acquisition. Foreign language annals. 2018;51(1):55-72. doi: 10.1111/flan.12314 521 7. Hamrick P, Lum JA, Ullman MT. Child first language and adult second language are both 522 tied to general-purpose learning systems. Proceedings of the National Academy of 523 Sciences. 2018;115(7):1487-92. 524 8. Lado B. Aptitude and pedagogical conditions in the early development of a nonprimary 525 language. Applied Psycholinguistics. 2017;38(3):679-701. 526 9. Faretta-Stutenberg M, Morgan-Short K. The interplay of individual differences and context 527 of learning in behavioral and neurocognitive second language development. Second 528 Language Research. 2018;34 (1): 67-101. doi: 10.1177/0267658316684903. 529 10. Tagarelli KM, Ruiz S, Moreno Vega JL, Rebuschat P. Variability in second language 530 learning: The roles of individual differences, learning conditions, and linguistic 531 complexity. Studies in Second Language Acquisition. 2016;38(2):293-316. doi: 532 10.1017/S0272263116000036.
- 533 11. Buffington J, Morgan-Short K. Declarative and procedural memory as individual
 534 differences in second language aptitude. In: Wen Z, Skehan P, Biedroń A, Li S, Sparks R,
 535 editors. Language aptitude: Multiple perspectives and emerging trends. New York:
 536 Routledge; 2019. pp. 215–237.
- 537 12. Marsden E, Morgan- Short K, Thompson S, Abugaber D. Replication in second language
 538 research: Narrative and systematic reviews and recommendations for the field. Language
 539 Learning. 2018;68(2): 321-91. doi:10.1111/lang.12286.

540	13. Plonsky L. Study quality in SLA: An assessment of designs, analyses, and reporting
541	practices in quantitative L2 research. Studies in Second Language Acquisition. 2013;35(4):
542	655-87. doi: 10.1017/S0272263113000399
543	14. Plonsky L. Quantitative research methods. In: Loewen S, Sato M, editors. The Routledge
544	handbook of instructed second language acquisition. New York: Routledge; 2017. pp. 505-
545	521.
546	15. Lindstromberg S. Inferential statistics in Language Teaching Research: A review and ways
547	forward. Language Teaching Research. 2016;20(6): 741-68. doi:
548	10.1177/1362168816649979.
549	16. Tackett JL, Brandes CM, King KM, Markon KE. Psychology's replication crisis and
550	clinical psychological science. Annual review of clinical psychology. 2019;15:579-604.
551	17. Krantz JH, Reips UD. The state of web-based research: A survey and call for inclusion in
552	curricula. Behavior Research Methods. 2017;49(5): 1621-1619. doi: 10.3758/s13428-017-
553	0882-x
554	18. Roever C. Web-based language testing. Language Learning & Technology. 2001;5(2):84-
555	94.
556	19. Domínguez C, López-Cuadrado J, Armendariz A, Jaime A, Heras J, Pérez TA. Exploring
557	the differences between low-stakes proctored and unproctored language testing using an
558	Internet-based application. Computer Assisted Language Learning. 2019:1-27.
559	20. Diaz Maggioli GH. Web- Based Testing. The TESOL Encyclopedia of English Language
560	Teaching. 2018:1-6.
561	21. Birnbaum MH. Human research and data collection via the Internet. Annu. Rev. Psychol
562	2004;55:803-32.

563	22. Hicks KL, Foster JL, Engle RW. Measuring working memory capacity on the web with
564	the online working memory lab (the OWL). Journal of Applied Research in Memory and
565	Cognition. 2016;5(4): 478-89. doi: 10.1016/j.jarmac.2016.07.010.
566	23. Wolfe CR. Twenty years of Internet-based research at SCiP: A discussion of surviving
567	concepts and new methodologies. Behavior research methods. 2017;49(5): 1615-1620. doi:
568	10.3758/s13428-017-0858-x.
569	24. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil
570	administration of patient-reported outcome measures: a meta-analytic review. Value in
571	Health. 2008;11(2): 322–333. doi: 10.1111/j.1524-4733.2007.00231.x.
572	25. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power
573	failure: why small sample size undermines the reliability of neuroscience. Nature Reviews
574	Neuroscience. 2013;14(5):365.
575	26. Branch MN. The "Reproducibility Crisis:" Might the Methods Used Frequently in
576	Behavior-Analysis Research Help?. Perspectives on Behavior Science. 2019;42(1):77-89.
577	27. Laraway S, Snycerski S, Pradhan S, Huitema BE. An overview of scientific
578	reproducibility: Consideration of relevant issues for behavior science/analysis.
579	Perspectives on Behavior Science. 2019;42(1):33-57.
580	28. Shrout PE, Rodgers JL. Psychology, science, and knowledge construction: Broadening
581	perspectives from the replication crisis. Annual review of psychology. 2018;69:487-510.
582	29. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ:
583	Erlbaum; 2013.
584	30. Stewart N, Chandler J, Paolacci G. Crowdsourcing samples in cognitive science. Trends in
585	cognitive sciences. 2017;21(10):736-48

587 Perspectives on Psychological Science. 2016;11(2): 239-64. doi:
588 10.1177/1745691615621279

31. Cowan N. Working memory maturation: Can we get at the essence of cognitive growth?.

- 589 32. Baddeley AD. Modularity, working memory and language acquisition. Second Language
 590 Research. 2017;33(3): 299-311. doi: 10.1177/0267658317709852.
- 33. Roehr K. Linguistic and metalinguistic categories in second language learning. Cognitive
 Linguistics. 2008;19(1): 67-106. doi: 10.1515/COG.2008.005.
- 593 34. Grundy JG, Timmer K. Bilingualism and working memory capacity: A comprehensive
 594 meta-analysis. Second Language Research. 2017;33(3): 325-40. doi:
 595 10.1177/0267658316678286.
- 596 35. Jeon EH, Yamashita J. L2 reading comprehension and its correlates: A meta- analysis.
 597 Language Learning. 2014;64(1):160-212. doi: 10.1111/lang.12034.
- 598 36. Linck JA, Osthus P, Koeth JT, Bunting MF. Working memory and second language
 599 comprehension and production: A meta-analysis. Psychonomic Bulletin & Review.
 600 2014;21(4): 861-83. doi: 10.3758/s13423-013-0565-2.
- 37. Bailey H, Dunlosky J, Kane MJ. Contribution of strategy use to performance on complex
 and simple span tasks. Memory & cognition. 2011;39(3): 447-61. doi: 10.3758/s13421010-0034-3.
- 38. Turner ML, Engle RW. Is working memory capacity task dependent?. Journal of memory
 and language. 1989;28(2): 127-54. doi: 10.1016/0749-596X(89)90040-5.
- 606 39. Conway ARA, Kane , MJ , Bunting MF, Hambrick DZ, Wilhelm O, et al. (2005) Working
 607 memory span tasks: A methodological review and user's guide. Psychonomic Bulletin and
 608 Review 12(12): 769–786. doi: 10.3758/BF03196772.

- 40. Zhou H, Rossi S, Chen B. Effects of working memory capacity and tasks in processing L2
 complex sentence: evidence from Chinese-English bilinguals. Frontiers in psychology.
 2017;8: 595. doi: 10.3389/fpsyg.2017.00595
- 612 41. Reber PJ, Knowlton BJ, Squire LR. Dissociable properties of memory systems: differences
- 613 in the flexibility of declarative and nondeclarative knowledge. Behavioral Neuroscience.
- 614 1996;110(5): 861. doi: 10.1037/0735-7044.110.5.861.
- 42. Squire LR. Memory systems of the brain: a brief history and current perspective.
 Neurobiology of learning and memory. 2004;82(3): 171-7. doi: 10.1016/j.nlm.2004.06.005
- 617 43. Eichenbaum H. Hippocampus: cognitive processes and neural representations that underlie
 618 declarative memory. Neuron. 2004;44(1):109-20.
- 44. Squire LR. Memory systems of the brain: a brief history and current perspective.
 Neurobiology of learning and memory. 2004;82(3): 171-7. doi:
 10.1016/j.nlm.2004.06.005.
- 45. Knowlton BJ, Siegel AL, Moody TD. Procedural learning in humans. In Byrne JH, editor.
 Learning and memory: A comprehensive reference. 2nd ed. Oxford: Academic Press;
 2017. pp. 295–312.
- 46. Hamrick P, Lum JA, Ullman MT. Child first language and adult second language are both
 tied to general-purpose learning systems. Proceedings of the National Academy of
 Sciences. 2018;115(7): 1487-1492. doi: 10.1073/pnas.1713975115.
- 47. Ullman MT. The declarative/procedural model: A neurobiologically motivated theory of
 first and second language. In: VanPatten B, Williams J, editors. Theories in second
 language acquisition: An introduction. 2nd ed. New York: Routledge; 2015. pp. 135-158.

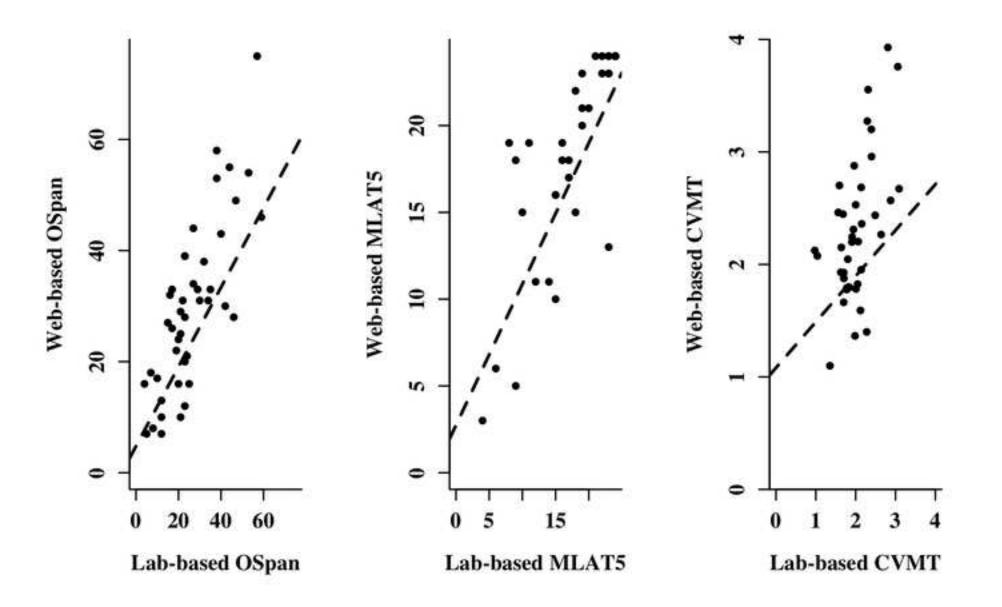
- 48. Ullman MT. The declarative/procedural model: A neurobiological model of language
 learning, knowledge, and use. In: Hickok G, Small SA, editors. Neurobiology of language.
 Amsterdam: Elsevier; 2016. pp. 498–505.
- 49. Morgan-Short K, Faretta-Stutenberg M, Brill-Schuetz KA, Carpenter H, Wong PC.
 Declarative and procedural memory as individual differences in second language
 acquisition. Bilingualism: Language and Cognition. 2014;17(1):56-72. doi:
 10.1017/S1366728912000715.
- 50. Carpenter, HS. A behavioral and electrophysiological investigation of different aptitudes
 for L2 grammar in learners equated for proficiency level. Ph.D. Thesis, Georgetown
 University. 2008. Available from: http://hdl.handle.net/10822/558127.
- 51. Carroll JB, Sapon SM. Modern Language Aptitude Test: Manual. New York:
 Psychological Corporation; 1959.
- 52. Trahan DE, Larrabee GJ. Continuous visual memory test. Odessa, FL: Assessment
 Resources. 1988.
- 53. Schneider W, Eschman A, Zuccolotto A. EPrime user's guide. Pittsburgh, PA: Psychology
 Software Tools Inc. 2002.
- 54. Unsworth N, Heitz RP, Schrock JC, Engle RW. An automated version of the operation
 span task. Behavior Research Methods. 2005;37(3): 498-505. doi: 10.3758/BF03192720.
- 55. Wickens TD. Elementary signal detection theory. New York: Oxford University Press;
 2002.
- 56. R Development Core Team (2016) R: A language and environment for statistical
 computing. R Foundation for Statistical Computing, Vienna,
 Austria. http://www.rproject.org.

- 57. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4.
 Journal of Statistical Software. 2015;67(1), 1–48. doi: 10.18637/jss.v067.i01
- 58. Kane MJ, Hambrick DZ, Tuholski SW, Wilhelm O, Payne TW, Engle RW. The generality

of working memory capacity: a latent-variable approach to verbal and visuospatial memory

- span and reasoning. Journal of Experimental Psychology: General. 2004;133(2): 189-217.
- 659 doi: 10.1037/0096-3445.133.2.189.
- 59. Gelman A. The failure of null hypothesis significance testing when studying incremental
 changes, and what to do about it. Personality and Social Psychology Bulletin. 2018;44(1):
 16-23. doi: 10.1177/0146167217729162.
- 663 60. Leidheiser W, Branyon J, Baldwin N, Pak R, McLaughlin A. Lessons learned in adapting
 664 a lab-based measure of working memory capacity for the web. In: Proceedings of the
 665 Human Factors and Ergonomics Society Annual Meeting 2015. Los Angeles: Sage CA;
 666 2015. pp. 756-760.
- 667 61. Reips UD, Krantz JH. Conducting true experiments on the Web. In: Gosling SD, Johnson
 668 JA, editors. Advanced methods for conducting online behavioral research. Washington,
- 669DC: American Psychological Association; 2010. pp. 193-216.
- 670 62. MacWhinney B. A shared platform for studying second language acquisition. Language
 671 Learning. 2017;67(S1): 254-75. doi: 10.1111/lang.12220.
- 672 63. Meurers D, Dickinson M. Evidence and interpretation in language learning research:
 673 Opportunities for collaboration with computational linguistics. Language Learning.
 674 2017;67(S1): 66-95. doi: 10.1111/lang.12233.
- 675 64. Ziegler N, Meurers D, Rebuschat P, Ruiz S, Moreno- Vega JL, Chinkina M, Li W, Grey
- 676 S. Interdisciplinary research at the intersection of CALL, NLP, and SLA: Methodological

- 677 implications from an input enhancement project. Language Learning. 2017;67(S1): 209-
- 678 231. doi: 10.1111/lang.12227.



1	Measuring individual differences in cognitive abilities in the lab and on the web
2	
3	
4	Simón Ruiz ^{1*} , Xiaobin Chen ² Chen ¹ , Patrick Rebuschat ^{1,3} - ² Detmar Meurers ^{1,43}
5	
6	
7	
8	
9	¹ [¶] LEAD Graduate School and Research Network, University of Tübingen, Tübingen, Germany
10 11	² . [¶] Department of Theoretical and Applied Linguistics, University of Cambridge, United Kingdom
11	- Department of Theoretical and Applied Eniguistics, Oniversity of Camonage, Onited Knigdom
12	³² [¶] Department of Linguistics and English Language, Lancaster University, Lancaster, United
14	Kingdom
15	
16	⁴ Department- <u>³Department</u> of Linguistics, University of Tübingen, Tübingen, Germany
17	
18	* Corresponding author
19	E-mail: simon.ruiz-hernandez@sfs.uni-tuebingen.de (SR)
20	
21	
22	[¶] These authors contributed equally to this work.
23	
24	

25 Abstract

26 The present study compared lab-based and web-based versions of cognitive individual 27 difference measures widely used in second language research (working memory and declarative 28 memory). Our objective was to validate web-based versions of these tests for future research and 29 to make these measures available for the wider second language research community, thus 30 contributing to the study of individual differences in language learning. The establishment of 31 measurement equivalence of the two administration modes is important because web-based testing 32 allows researchers to address methodological challenges such as restricted population sampling, 33 low statistical power, and small sample sizes. Our results indicate that the lab-based and web-34 based versions of the tests were equivalent, i.e., scores of the two test modes correlated. The 35 strength of the relationships, however, varied as a function of the kind of measure, with 36 equivalence appearing to be stronger in both the working memory and the verbal declarative 37 memory tests, and less so in the nonverbal declarative memory test. Overall, the study provides 38 evidence that web-based testing of cognitive abilities can produce similar performance scores as 39 in the lab.

40

41 Introduction

Individual differences can greatly affect how we acquire and process language [1-3] and mediate and/moderate the effectiveness of instruction [4]. In adult language learning, for example, learners' cognitive abilities have great explanatory power in accounting for differences in learning outcomes ([5-6]). For instance, Among these, working memory and declarative memory are considered to be particularly important sources of learner variation (e.g., [7-10]; see [4, 11], for reviews). 48 The effect of working memory and declarative memory on language learning has been 49 primarily studied in lab settings, i.e., in well-controlled environments where participants are tested 50 individually. While this choice is methodologically sound, it can also negatively affect sample size 51 and population sampling [12, 13, 143, 14]. Lab-based testing generally means testing participants 52 individually and sequentially, which is labor-intensive and could explain why lab studies tend to 53 have (too) few participants to allow for meaningful generalization. For exampleAs a way of an 54 example, in second language (L2) research, Plonsky [13] found that the typical sample size in L2 55 studies was 19 participants, and Lindstromberg [15] recently reported a similar small average 56 sample size of 20 participants. In the same vein, [16] reported that, in psychology, median sample 57 sizes have not increased considerably in the last two decades, and are generally too small to detect small effect sizes, which are distinctive of many psychological effects. Moreover, many (if not 58 59 most) lab studies in $\frac{1}{1}$ -research draw their sample from the surrounding student population, which 60 is understandable given the ease of access, but also means that samples are often not representative 61 of the population of interest. Conducting research by means of remote testing via the web could 62 alleviate some of these concerns. For example, web-based testing facilitates the acquisition of large 63 amounts of data since participants can be tested simultaneously, which in turn enablinges 64 researchers to run higherer-powered studies. Likewise, test administration can also be more cost-65 effective than research conducted in the lab [1517]. Web based experimenting has been found to be a reliable and effective research tool [16,17, 18]. 66 67

<u>The use of (remote) web-based testing can also offer other important methodological</u>
 <u>advantages over other forms of simultaneous delivery of tests, such as traditional paper-pencil and</u>
 (offline) computer-based testing [18, 19]. Particularly, it allows researchers to standardize and

optimize testing procedures, which can contribute to more consistent and uniform test-taking conditions across different locations and times [20]. This, in turn, can also facilitate the replication of studies [21]. Moreover, remote testing via the web can too-reduce experimenter effects, as testing can occur in more ecologically-valid settings, and without any direct contact between experimenters and participants [20, 21]. Finally, and more importantly, web-based experimenting has been found to be a reliable and effective research tool [17, 22, 23].

77 The present study compared lab-based and web-based versions of cognitive tests that are 78 widely used in disciplines such as psychology and second language research. Particularly, our The 79 intent was to compare performance of measures as they are originally used in the lab with their 80 corresponding online versions. In doing so, our objective was to validate the web-based tests for 81 use in subsequent research and to make these available to the wider second language research 82 community, and especially to researchers working on the area of L2 acquisition. T. The sharing of 83 tasks, especially of tasks that permit the collection of substantial amounts of data via the web, will 84 be an important component in reducing alleviating the data collection issues problem associated 85 with lab-based research in SLA. Moreover, making these specific tasks available will also 86 contribute directly to our understanding of individual differences in L2 acquisition. To support 87 such task sharing and use, it is essential to first establish the validity of the online versions of the 88 tasks (on a par with what is established about the offline versions). With this in mind, the study set 89 out to establish measurement equivalence between lab-based and web-based tests of working 90 memory and declarative memory.

According to Gwaltney, Shields and Shiffman ([<u>1924</u>], p. 323), measurement equivalence can be established if "1) the rank orders of scores of individuals tested in alternative modes closely approximate each other; and 2) the means, dispersions, and shapes of the score distributions are

94 approximately the same". The first type of equivalence is reregards lated to whether differences 95 observed found-in one measurement are also systematically found found-in the other, meaning 96 that, .- This means that, even when although the two measurements produce estimate two different 97 numbers, these numbers are clearly and have a systematically and very clear relationship 98 associated withto each other. The second type concerns whether two measurements yield the same 99 numbers. Considering that this study was a piecework is a subcomponent of the dissertation 100 research of the first author, with limitinged funding and time (see limitations below), it was 101 therefore decided to undertake a morewe focused the investigation on by looking at only one type 102 of measurement equivalence, in this case, the first type: Do people who have relatively high values 103 in one of tests also have relatively high values on the other test, and the other way around? More 104 specifically, we compare the differential performance generated by two versions of tests measuring 105 working memory and declarative memory abilities-capacities in lab-based and web-based settings, 106 in order with the aim to determinassesse whether the two versions are equivalent regarding with 107 respect to the relationships between scores.

108 Assessing measurement equivalence between these two administration modes lab and web-109 based measurements is essential for several reasons. Firstly, it is necessary to demonstrate show 110 that the findingsresults obtained in of web-based studies are comparable to those of previous 111 research, which have been mainly collected e predominantly obtained from data gathered in lab-112 based settings. Secondly, it is imperative important to ensure to ensure that cognitive constructs 113 constructs are similarly gauged measured in the same way in both testing modalities modes. 114 FinallyLikewise, it is important-crucial to ascertain-establish whether lab-based and web-based tests measures are equivalent, given that because, if they are, web-based testing could prove be a 115 feasible alternative to be a viable way to tackle address some of the current methodological 116

117	_issues found in L2 research conducted in lab-based settings, such as underpowered studies,
118	restricted population sampling, and small sample sizes, among others-[1317, 1422, 23]. Of these
119	methodological issues, in particular, low statistical power and small sample sizes have been
120	identified as key factors in the ongoing discussions about the reproducibility of research findings
121	in life and social sciences [25-27]. In psychology, for example, there is currently considerable
122	debate about the so-called replication crisis [28], that is, failure to reproduce significant findings
123	when replicating previous research [27]. In this regard, and considering that much research is
124	underpowered [29, 30], web-based testing can enable the collection of larger sample sizes, and
125	thus contribute to achieve more statistical power to detect the effects of interest. On the other hand,
126	the ease of access, cost-effectiveness, and practicality of web-testing can also increase the attempts
127	to reproduce results from previous studies, and thus making (large-scale) replication studies more
128	appealing for researchers to undertake [30].
1	

- 129
- 130

131 Working memory

Working memory <u>is_refers_to</u>_the capacity to <u>simultaneously</u>_process and <u>hold_retain</u> information <u>at the same time_while performing_earrying_out</u>_complex_cognitive tasks such as language learning, comprehension and production [2031]. <u>According to Following</u> Baddeley and colleagues (e.g., [2132]), working memory is a multicomponent system that <u>includes_consists of</u> storage subsystems that are-responsible for <u>retaining both_holding</u>_visual-spatial and auditory information, an episodic buffer that <u>serves acts</u> as a link between the storage subsystems and longterm memory, and a central executive that <u>actsfunctions</u> as an attentional control system. 139 Regarding In-L2 learning, working memory appears to assists learners to simultaneously 140 jointly process form, meaning and use of language forms at the same time. More specifically, 141 working memory is involved in key cognitive processes such as decision making, attention control, 142 explicit deduction, information retrieval and analogical reasoning [4]. Moreover, working memory 143 is also important for retaining metalinguistic information while comprehending and producing L2 144 language [2233]. In this regard, meta-analytic work has reported the important role of working 145 memory in L2 comprehension and production (e.g., [2334-2536]). For example, Linck et al. 146 ([2536], p. 873) found that working memory has a positive impact on L2 comprehension outcomes 147 (r = 0.24). Likewise, Jeon and Yamashita's [2435] meta-analysis also showed that working 148 memory is related to L2 reading comprehension (r = 0.42). Regarding production, meta-analytic 149 research has, too, indicated a significant association with working memory (e.g., [2536]). In this 150 case, Linck et al. ([2536], p. 873) found a positive correlation for productive outcomes as well (r 151 = 0.27).

152 Working memory is often measured by means of simple or complex span tasks. Simple 153 span tasks, such as (e.g., digit span and letter span,) entails involve recalling short lists of items, 154 and they seek to gauge-measure the storage component aspect of working memory [2637]. 155 Complex span tasks, such as the operation span task (OSpan; [2738]), on the other hand, entail 156 include remembering stimuli while performing a another secondary task. This type of tasks, and 157 are thought to taxtaxes both processing (attention) and storage (memory) aspects of components 158 of working memory [2132]. Here, we focus on a complex task, namely the OSpan. This complex 159 task has been found to be a valid and reliable measure of working memory capacity [2839], and 160 has also been recommended as a more accurate measure to examine the association between 161 working memory and L2 processing and learning [2940].

162

163 **Declarative memory**

Declarative memory is the capacity to consciously recall and use information [3041]. The declarative memory system is one of the long-term memory systems in the brain [3142]. It is mainly responsible for the processing, storage, and retrieval of information about facts (semantic knowledge) and events (episodic knowledge; [3243, 4343]). Learning in the declarative memory system is quick, intentional, and attention-driven [4534].

169 Substantial research has now investigated the role of declarative memory in first and 170 second language acquisition [3546]. In first language acquisition, declarative memory is appears 171 to be involved in the processing, storage and learning of both arbitrary linguistic knowledge (e.g., 172 word meanings) as well as rule-governed aspects of language (e.g., generalizing grammar rules 173 [3647, 3748]). In the case of L2 acquisition, declarative memory appears to underpins the learning, 174 storage and processing of L2 vocabulary and grammar $[\frac{47, 4836, 37}{1000}]$, at least in the earliest phases 175 of acquisition [3546, 3849]. Several studies (e.g., [2, 9, 3849, 5039]) has confirmed the predictive 176 ability of declarative memory to explain variation in L2 attainment.

Declarative memory has been tested through recall and recognition tasks (e.g., <u>3849</u>, <u>5039</u>),
both verbal, such as the paired associates subtest of the Modern Language Aptitude Test (MLAT5;
[4051]), and nonverbal, such as the Continuous Visual Memory Task (CVMT; [4152]).

180

181 The present study

182 The main goal of the present study was to provide web-based versions of commonly employed 183 individual difference measures in second language research, in order to make them usable in large-184 scale intervention studies (generally in authentic, real-life learning contexts). To that end, we examined whether lab-based and web-based versions of working memory and declarative memory tests yield similar performance scores, i.e., whether the two versions were equivalent or comparable. More specifically, we assessed whether the values of one type of mode of administration corresponded to the values in the other mode (i.e., first type of equivalence). In other words, are the differences in scores constant, or parallel in the two ways of measuring? The web-based versions are freely available; to use the test, please send an email to the first author.

191

192 Methods

193 **Ethics statement**

This research was approved by the Commission for Ethics in Psychological Research, University of Tübingen, and all participants provided written informed consent prior to commencement of the study.

197

198 **Participants**

199 Fifty participants (37 women and 13 men), with a mean age of 26.4 years (SD = 4.2), 200 partook part in the study. Most The majority of participants were native speakers of German (72%), 201 followed by Russian (8%), Spanish (6%), Chinese (4%), English, Hungarian, Persian, Serbian and 202 Vietnamese (2% each). Seven (14%) participants did not complete the second half of the study 203 (i.e., web-based testing). Additionally, participant numbers differed across test versions due to 204 technical difficulties (i.e., participants erroneously entered their responses using the 205 keyboardwrong keys [Web-based CVMT]; and data was missingnot correctly saved for one 206 participant [Web-based MLAT5]; see description and Results; Table 1 below, and Discussion).-207 Twenty-seven participants were graduate students (54%), and twenty-three were undergraduates

208 (46%). Participants self-reported English proficiency, with most being advanced learners (82%),
209 followed by intermediate (18%). All subjects gave informed consent and received €20 for
210 participating.

211

212 Materials

213 Three cognitive tests were administered, one assessing measuring working memory 214 capacity, and two indexing assessing verbal and nonverbal declarative memory capacity abilities, 215 respectively. In the lab-based contexsettingt, both working memory and nonverbal declarative 216 memory tests were programmed and delivered via E-Prime v2.0 [4253]; the verbal declarative 217 memory test was givenapplied in paper-pencil form, as originally developed and delivered. 218 Moreover, For the wweb-based mode, versions of the three cognitive tests were developed for 219 this study using Java with the GoogleWeb Toolkit (http://www.gwtproject.org), and were 220 accessible from all browsers. A description of each The tests is given below. are described below. 221

222 Working memory. To assess participants' working memory capacity, aAn adapted version of 223 the Automated Operation Span Task (OSpan; [4354]), a computerized form of the complex span 224 task created by Turner and Engle [2738], was used to gauge participants' working memory 225 capacity [9, 2217]. This adaptation was bBased on the Klingon Span Task implemented 226 developed by Hicks et al. [1722], this version and consisted of using Klingon symbols replacing 227 instead of letters, -(the original stimuli to be remembered in the original OSpan task) with 228 Klingon symbols. In Hicks et al.' study, participants cheated by writing down the letter 229 memoranda in the web-based version of the classic OSpan, causing motivating Hicks et al. 230 implemented thehis change change of the original stimuli. because their research showed that

participants were cheating by writing down the letter memoranda in the web-based version of the
 classic OSpan.

233 The task took approximately 25 minutes to complete, 234 and included was divided into a practice phase and a testing phase. In the practice phase, 235 participants were first presented shown with a series of Klingon symbols on the screen, and then 236 were asked to recallemember them in the order in which they had appeared after at the end of 237 each trial (i.e., symbol recall). Next, participants were requiredasked to solve a series of simple 238 math operations equations (e.g., 85 * 42 + 71 = ?). Finally, subjects performed the symbol recall 239 while also solving the math problems, as they would do later do in the actual testing phase. 240 Following After the practice phase, participants were presented shown with the real trials, which 241 consisted of a list of 15 sets of 3–7 randomized symbols that appeared intermingled mixed with 242 the equations. In sum, there were, totaling 75 symbols and 75 math problems. At the end of each 243 set, participants were asked to remembereall the symbols in the sequence they had been 244 presented shown. An individual time limit to answer the math problems in the real trials was 245 calculated derived from the average response time plus 2.5 standard deviations taken during the 246 math practice section. Following Unsworth et al. [4654], a partial score (i.e., total number of 247 correct symbols recalled in the correct order) was taken as the OSpan score (see [2839], for a 248 description of scoring procedures). The highest possible score was 75. The entire task took about 249 <u>25 min.</u>

250

Verbal declarative memory. <u>To measure verbal declarative memory, t</u>The Modern
 Language Aptitude Test, Part 5, Paired Associates (MLAT5; [4051]), was used as a verbal
 measure of declarative memory [9, 3849, 5039]. <u>In t</u>The MLAT5, <u>required participants were</u>

254 required to memorize artificial, pseudo-Kurdish words and their meanings in English. 255 Participants were first asked to study ied-24-word association pairs for two minutes, and then 256 completeed a two-minute practice section. During the practice section, Tthe list of foreign words 257 with and their respective English equivalents meanings was were made available for participants 258 as they completed the practice sessionto refer back if they needed to. Finally, subjects were 259 instructed to completed a timed multiple-choice test (four minutes), by in which they were asked 260 to-selecting the English meaning of each of the 24 pseudo-Kurdish words from five options 261 previously seen-displayed at the memorization stage. For each correct response, one point was 262 awardedgiven, yielding a total score of 24 points. The test duration was about 8 minutes.

263

264 **Nonverbal declarative memory.** The Continuous Visual Memory Task (CVMT; [4652]) 265 served was included as a measure n assessment of nonverbal declarative memory [9, 3849, 50] 266 39]. As a visual recognition test, -tThe CVMT is a visual recognition test that involves entails 267 asking participants to first view a collection of complex abstract designs on the screen, and then 268 to indicate whether the image they just saw was novel ("new") in the collection, or they had seen 269 the image before ("old"). Seven of the designs were "old" (target items), and 63 were "new" 270 (distractors). Throughout the task, tThe target items appeared seven times (49 trials), and the 271 distractors only once (63 trials) across the test. All items were presented shown in a random but 272 fixed order, each one appearing on the screen for two seconds. After-Following the two seconds, 273 participants were instructed to respond to the "OLD or NEW?" prompt on the screen. In the lab-274 based settingmode, subjects used mouse click for making indicated their choice, by mouse 275 elicking either left for "NEW", or right for "OLD". In the web-based-modesetting, they 276 responded by pressing either the "N" key for "NEW", or the "O" key for "OLD" on the

keyboard.<u>-TOverall, the CVMT took required</u> 10 minutes to be completed. For each participant,
a<u>A</u> d'(d-prime) score [44<u>55</u>] for CVMT was calculated for each participantomputed. The d'
score was used to account forreduce potential the possible participants' response bias toward
choosing "OLD" or "NEW".

- 281
- 282

283 **Procedure**

As previously noted, <u>As previously noted</u>, <u>pp</u>articipants <u>completed underwent</u> two cognitive testing sessions, one in the lab and one on the web. In the lab-based session, <u>with the</u> assistance of a proctor in the presence of a proctor, each subject was tested individually. After providing informed consent, participants took the three cognitive tests under investigation in fixed order: OSpan, CVMT, and MLAT5. <u>Upon finishing the MLAT5</u>, <u>subjects then They were then</u> asked to filled <u>out in a</u> background questionnaire. The whole lab-based session <u>lasted took</u> about 40 mi<u>nnutes</u>.

291 For-Regarding the web-based session, each subject was sent an email with -containing a 292 unique web link with a personalized code, that when which once clicked, took them to an interface 293 that hosted housing the web-based versions of the cognitive tests. In order tTo avoid prevent 294 participants from taking the tests-multiple- responses by the same participanttimes, the link was 295 disabled became nonfunctional once subjects they had submitted their responses in the last test 296 (i.e., MLAT5). In the email, participants were also informed that the web-based session lasted 297 about 40 minutes, and that it had to be completed within a week. On the interface, following 298 informed consent, subjects were given provided with general instructions that reflected in 299 accordance with the nature of a the web-based nature of the experiment. These Such instructions

included completing the experiment in a quiet place without interruption, and from start to finish in one sitting. Likewise, the Participants were also instructed not to-use of the browser's back button, or-refreshing the browser page, or close closing the browser window were prohibited. Importantly, they participants were told instructed not to take any notes during theat any point during the entire experiment. The web-based tests were taken given in the same fixed order as in the lab-based session. On average, tThe mean period between the first and second testing was 45.7 days (SD = 4.1).

307

308 **Results**

309 All data were analyzed using by means of the statistical software package R (version 3.3.2; 310 ([56] R Core Team, 2016). Missing data was ignored (complete--case analysis). Linear regression 311 models were built using the lm function in the lme4 library [57]. From a temporal 312 perspective point of view, lab scores were used to predict web scores in the linear regression 313 models. To verify normality, model residuals were visually inspected. Reliability was assessed 314 using Cronbach's alpha. Following Kane et al. [4558], for the lab-based working memory test 315 (OSpan-Lab-based), reliability was assessed by calculating the proportion of correctly recalled 316 Klingon symbols per each of the 15 trials in the test (e.g., one out of four symbols correctly 317 recalled corresponded to a proportion of .25). For the web-based working memory test (OSpan-318 Web-based), however, internal consistency is not reported, since it was not technically possible 319 to perform a detailed item-based analysis. Descriptive statistics are presented first, followed by 320 correlations, internal consistency estimates (Cronbach's alpha), and the results of linear 321 regression analyses.

323 **Descriptive statistics**

Table 1 presents the descriptive statistics <u>for summarizing participants' participants'</u> performance on the three cognitive tests <u>under investigation</u> in both test<u>ing-settingsmodes</u>.

328 Table 1. Descriptive statistics for comparison of lab-based and web-based testing.

Test	Ν	М	SD	Skew	Kurtosis
OSpan Lab-based	50	25.78	13.34	0.61	2.90
OSpan Web-based	43	29.79	15.42	0.67	3.26
MLAT5 Lab-based	50	17.92	5.50	-0.64	2.49
MLAT5 Web-based	42	19.10	5.81	-1.19	3.58
CVMT Lab-based	49	1.99	0.46	0.23	3.35
CVMT Web-based	40	2.30	0.63	0.73	3.32

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task.

329

330 Correlations

331 Table 2 and Fig 1 show the correlations between/among the different versions of the

332 individual difference tests.

333

335 Table 2. Correlations between lab-based and web-based scores for individual difference

336 **tests.**

Test	OSpan	OSpan	MLAT5	MLAT5	CVMT
	Lab-based	Web-based	Lab-based	Web-based	Lab-based
OSpan Web-based	.80				
MLAT5 Lab-based	.40	.51			
MLAT5 Web-based	.32	.40	.82		
CVMT Lab-based	.19	.31	.42	.23	
CVMT Web-based	.21	.30	.21	.19	.55

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task.

337

338 Fig 1. Scatterplots of the correlation of each pair of lab-based and web-based versions of

- 339 individual difference measures.
- 340 OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern
- 341 Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous
- 342 Visual Memory Task.
- 343

344 **Reliability**

- 345 Table 3 presents Cronbach's alpha values of individual test versions.
- 346

347

Test	Cronbach's alpha		
OSpan Lab-based	.86		
MLAT5 Lab-based	.77		
MLAT5 Web-based	.93		
CVMT Lab-based	.63		
CVMT Web-based	.67		

Table 3. Cronbach's alphas for cognitive test versions.

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task.

349

350 Regression analysis

351 The results of the regression analyses are displayed in Table 4. For the working memory

test (OSpan), the unstandardized coefficient was .89 ($\beta = .77$, SE = 0.10, p < .001). For the verbal

declarative memory test (MLAT5), the unstandardized coefficient was .83 (β = .78, SE = 0.09, p

354 <.001). And for the nonverbal declarative memory test (CVMT), the unstandardized coefficient

355 was .74 (β = .54, SE = 0.19, p < .001). Overall, the results indicated that the lab-based and web-

based scores are substantially related.

357

Test	Unstandardized coefficient ^a	SE	р
OSpan	0.89 (.77)	0.10	<.001
MLAT5	0.83 (.78)	0.09	< .001
CVMT	0.74 (.54)	0.19	< .001

359 Table 4. Regression for comparison of lab-based and web-based scores.

Note: OSpan = Automated Operation Span Task; Verbal declarative memory test: MLAT5 = Modern Language Aptitude Test, Part 5; Nonverbal declarative memory test: CVMT = Continuous Visual Memory Task. ^aThe standardized coefficient (β) in parentheses.

360

361 **Discussion**

362 Studies on individual differences in language learning frequently assess the working 363 memory and declarative memory capacities of their participants in order to determine the effect 364 of these cognitive variables on learning outcomes. Most of this research, however, is conducted 365 in lab-based settings, which often implies relatively small sample size and a restricted population 366 sampling. Both of these methodological challenges can be addressed by means of remote testing 367 via the web. In the present study, we compared lab-based and web-based individual difference 368 measures in order to validate web-based tests for future research. The type of comparison 369 contributes significantly to ongoing efforts to improve the methodological robustness of current 370 second language research, for example [4712]. If web-based testing can be shown to yield 371 comparable results to lab-based testing, researchers will be able to reach more participants for 372 their studies, which, in turn, can help alleviate some of the current concerns in L2-lab-based 373 research (e.g., low statistical power, non-representative population samples, and small sample

sizes). In addition, demonstrating the equivalence of lab-based and web-based measures of the
same individual difference constructs is essential for the comparability of results across studies.
Crucially, establishing measurement equivalence between lab-based and web-based versions will
also provide assurance that the tests are measuring cognitive constructs the same way regardless
of administration mode [1617, 4859].

379 Findings The results showed indicated that the scores in the lab-based and web-based 380 versions of three cognitive tests (MLAT5, CVMT, OSpan) were equivalent in the sense 381 that concerning differences in performance, which were constant in the two versions, . This 382 suggestings that participants who had relatively high values in one task also had relatively high 383 values in the second, or the other way around. However, the strength of the association 384 relationship was a function of depended on the kind of test. More specifically, iIn both the 385 working memory test (OSpan) and the verbal declarative memory test (MLAT5), the scores were 386 more strongly correlated ($\beta = .77$ and $\beta = .78$, respectively); for the nonverbal declarative test 387 (CVMT), equivalence appears to be weaker ($\beta = .54$).-OverallOn the whole, the correlations 388 reported here between lab-based and web-based scores are consistent with the assumption that 389 both versions seem to likely measure the same cognitive construct, at least for the working 390 memory test (OSpan) and the verbal declarative memory test (MLAT5), and, to a lesser extent, 391 for the nonverbal declarative test (CVMT).

A possible-potential explanation for lesser the weaker equivalence in found for the versionss of the nonverbal declarative test (CVMT) is perhapscould be due to the different difference in the waymanner in which rthe responses to the visual stimuli were input entered in the two testing modes. It will be Recall recalled that in the lab-based version, participants used 396 left ("NEW") or right ("OLD") mouse clicking to provide aenter their response, whereas in the 397 web-based version, they used the keyboard ("N" and "O" keys). This is-modification made to 398 was made to the web-based version was motivated by because of technical reasons, specifically, 399 , i.e., the browser window may not register the participants' response if the cursor is not over a 400 certain area on the page, which in itself which in turn may may cause problems of missing data. 401 Previous research It has been previously reported has found that participants in web-based 402 research are particularly prone prone to err make errors when using the keyboard to input enter 403 their responses [4960], which in this case might have affected the results of the comparison 404 between lab-based and web-based versions of CVMT. Future research Further studies-comparing 405 performance between the two versions lab and web-versions may benefit from collecting 406 gathering data vithrougha touch input instead, as this which might help overcome potential the 407 technical difficulties y of caused by using employing mouse clicking for web-based data. 408 collection reported here. 409 **Limitations** 410 411 Some limitations of the study and the findings presented here should be considered. One 412 of the limitations was the small sample size. As mentioned earlier, logistic constrains due to the 413 availability of time and funding prevented the researchers from testing more participants for this

- 414 <u>study. In addition, the fact that some participants (14%) dropped out before completing any of</u>
- the web-based measures in the second part of the experiment, which is typical in web-based
- 416 research [17], also contributed to the reduction of the data available for the comparison between
- 417 <u>lab and web-based testing in the present investigation. Therefore, our findings need to should be</u>

418	replicated in a larger study. A second limitation was that test-retest reliability was not examined
419	here, given that the main aim of this study was to establish valid online versions of known
420	individual difference measures. Future research should assess test-retest reliability, as it is as an
421	interesting endeavor for studying individual difference measures in future work. Finally, and as
422	indicated above, a third limitation concerned technical issues that affected data collection, as
423	some participants used the wrong keys on the keyboard to submit their responses to the web-
424	based version of the CVMT, rendering the data from some of the participants impossible to use
425	for the comparison; furthermore, data from one subject was missing in the Web-based MLAT,
426	which could may have been due to technical problems originated from issues at the participant's
427	end (e.g., not following the general instructions given, such as refreshing or closing the browser
428	page [see Procedure]; or Internet disconnection). In this sense, Reips and Krantz [61] (see
429	also[17]) caution researchers that one of the potential disadvantages of Internet-driven testing is
430	the technical variability characteristic of web-based research (e.g., different browsers and
431	Internet connections), which, in turn, may affect data collection.
432	
433	Conclusion

<u>Despite the limitations, there are important contributions in this study.</u> This study aimed to establish the validity of using web-based versions of established offline tasks. As such, the study has provided evidence that it is <u>possible possible</u> to measure individual differences in cognitive abilities on the web and obtain similar performance as in the lab. The lab-based and web-based versions of the three cognitive tests are comparable or equivalent. However, given that they do not perfectly correlate, we recommend using one of the two modes within one study 440 and not comparing individual scores from one mode with scores from the other. Moreover, the 441 extent to which the measures are equivalent varies according to the test. In this sense, we are 442 confident that the two versions for the working memory test (OSpan) and the verbal declarative 443 memory (MLAT5) are fairly possibly measuring likely to measure the same construct, whereas 444 the correlation but we refrain from making such a strong statement between for the nonverbal 445 declarative test (CVMT) versions was less pronounced, where the two modes might still 446 plausibly measure strongly different aspects as well. Our research has shown that collecting 447 experimentally controlled data on cognitive individual differences typically used in <u>L2-the area</u> 448 of L2 research in the Internet is feasible and comparable to lab-based collection. Consequently, 449 some of these web-based versions could very well be incorporated, for example, in future web-450 based intervention studies on second language learning, thereby contributing to the scaling up of 451 data collection in the field $[\frac{5062-5264}{2}]$.

452

453 Acknowledgements

454

We would like to thank Johann Jacoby, for his invaluable for his valuable suggestions that strengthened our experimental design and analysis.

457

458 **References**

459

463	2. Hamrick P. Declarative and procedural memory abilities as individual differences in
464	incidental language learning. Learning and Individual Differences. 2015;44:9-15. doi:
465	10.1016/j.lindif.2015.10.003.
466	3. Ruiz S, Tagarelli KM, Rebuschat P. Simultaneous acquisition of words and syntax: Effects
467	of exposure condition and declarative memory. Frontiers in Psychology. 2018 12;9:1168.
468	doi: 10.3389/fpsyg.2018.01168
469	4. Li S. Cognitive differences and ISLA. In: Loewen S, Sato M, editors. The Routledge
470	handbook of instructed second language acquisition. New York: Routledge; 2017. pp. 396-
471	417.
472	4.5. Pawlak M. Overview of learner individual differences and their mediating effects on the
473	process and outcome of interaction. In Gurzynski-Weiss L., editor. Expanding individual
474	difference research in the interaction approach: Investigating learners, instructors, and
475	other interlocutors. Amsterdam: John Benjamins; 2017. pp. 19-40.
476	5.6. Larsen- Freeman D. Looking ahead: Future directions in, and future research into, second
477	language acquisition. Foreign language annals. 2018;51(1):55-72. doi: 10.1111/flan.12314
478	6.7. Hamrick P, Lum JA, Ullman MT. Child first language and adult second language are both
479	tied to general-purpose learning systems. Proceedings of the National Academy of
480	Sciences. 2018;115(7):1487-92.
481	7.8. Lado B. Aptitude and pedagogical conditions in the early development of a nonprimary
482	language. Applied Psycholinguistics. 2017;38(3):679-701.
483	8.9. Faretta-Stutenberg M, Morgan-Short K. The interplay of individual differences and context
484	of learning in behavioral and neurocognitive second language development. Second
485	Language Research. 2018;34 (1): 67-101. doi: 10.1177/0267658316684903.

- 486 9.10. Tagarelli KM, Ruiz S, Moreno Vega JL, Rebuschat P. Variability in second
 487 language learning: The roles of individual differences, learning conditions, and linguistic
 488 complexity. Studies in Second Language Acquisition. 2016;38(2):293-316. doi:
 489 10.1017/S0272263116000036.
- 490 <u>10.11.</u> Buffington J, Morgan-Short K. Declarative and procedural memory as individual
 491 differences in second language aptitude. In: Wen Z, Skehan P, Biedroń A, Li S, Sparks R,
 492 editors. Language aptitude: Multiple perspectives and emerging trends. New York:
 493 Routledge; 2019. pp. 215–237.
- 494 <u>11.12.</u> Marsden E, Morgan- Short K, Thompson S, Abugaber D. Replication in second
 495 language research: Narrative and systematic reviews and recommendations for the field.
 496 Language Learning. 2018;68(2): 321-91. doi:10.1111/lang.12286.
- 497 <u>12.13.</u> Plonsky L. Study quality in SLA: An assessment of designs, analyses, and reporting
 498 practices in quantitative L2 research. Studies in Second Language Acquisition. 2013;35(4):
 499 655-87. doi: 10.1017/S0272263113000399
- 500 <u>13.14.</u> Plonsky L. Quantitative research methods. In: Loewen S, Sato M, editors. The
 501 Routledge handbook of instructed second language acquisition. New York: Routledge;
 502 2017. pp. 505-521.
- 503 <u>15.</u> Lindstromberg S. Inferential statistics in Language Teaching Research: A review and ways
 504 forward. Language Teaching Research. 2016;20(6): 741-68. doi:
 505 10.1177/1362168816649979.
- 506 14.16. Tackett JL, Brandes CM, King KM, Markon KE. Psychology's replication crisis
 507 and clinical psychological science. Annual review of clinical psychology. 2019;15:579 508 <u>604.</u>

- 509 <u>15.17.</u> Krantz JH, Reips UD. The state of web-based research: A survey and call for
 510 inclusion in curricula. Behavior Research Methods. 2017;49(5): 1621-1619. doi:
 511 10.3758/s13428-017-0882-x
- 512 18. Roever C. Web-based language testing. Language Learning & Technology. 2001;5(2):84-
- 513 <u>94.</u>
- 514 <u>19. Domínguez C, López-Cuadrado J, Armendariz A, Jaime A, Heras J, Pérez TA. Exploring</u>
 515 <u>the differences between low-stakes proctored and unproctored language testing using an</u>
- 516 Internet-based application. Computer Assisted Language Learning. 2019:1-27.
- 517 <u>20. Diaz Maggioli GH. Web- Based Testing. The TESOL Encyclopedia of English Language</u>
 518 Teaching. 2018:1-6.
- 519 <u>21. Birnbaum MH. Human research and data collection via the Internet. Annu. Rev. Psychol.</u>
 520 2004;55:803-32.
- 521 <u>16.22.</u> Hicks KL, Foster JL, Engle RW. Measuring working memory capacity on the web
 522 with the online working memory lab (the OWL). Journal of Applied Research in Memory
 523 and Cognition. 2016;5(4): 478-89. doi: 10.1016/j.jarmac.2016.07.010.
- 524 17.23. Wolfe CR. Twenty years of Internet-based research at SCiP: A discussion of
 525 surviving concepts and new methodologies. Behavior research methods. 2017;49(5): 1615526 1620. doi: 10.3758/s13428-017-0858-x.
- 527 <u>24.</u> Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil
 528 administration of patient-reported outcome measures: a meta-analytic review. Value in
 529 Health. 2008;11(2): 322–333. doi: 10.1111/j.1524-4733.2007.00231.x.

530	25. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power
531	failure: why small sample size undermines the reliability of neuroscience. Nature Reviews
532	Neuroscience. 2013;14(5):365.
533	26. Branch MN. The "Reproducibility Crisis:" Might the Methods Used Frequently in
534	Behavior-Analysis Research Help?. Perspectives on Behavior Science. 2019;42(1):77-89.
535	27. Laraway S, Snycerski S, Pradhan S, Huitema BE. An overview of scientific
536	reproducibility: Consideration of relevant issues for behavior science/analysis.
537	Perspectives on Behavior Science. 2019;42(1):33-57.
538	28. Shrout PE, Rodgers JL. Psychology, science, and knowledge construction: Broadening
539	perspectives from the replication crisis. Annual review of psychology. 2018;69:487-510.
540	29. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ:
541	<u>Erlbaum; 2013.</u>
542	18.30. Stewart N, Chandler J, Paolacci G. Crowdsourcing samples in cognitive science.
543	Trends in cognitive sciences. 2017;21(10):736-48
544	<u>19.31.</u> Cowan N. Working memory maturation: Can we get at the essence of cognitive
545	growth?. Perspectives on Psychological Science. 2016;11(2): 239-64. doi:
546	10.1177/1745691615621279
547	20.32. Baddeley AD. Modularity, working memory and language acquisition. Second
548	Language Research. 2017;33(3): 299-311. doi: 10.1177/0267658317709852.
549	21.33. Roehr K. Linguistic and metalinguistic categories in second language learning.
550	Cognitive Linguistics. 2008;19(1): 67-106. doi: 10.1515/COG.2008.005.

- 551 <u>22.34.</u> Grundy JG, Timmer K. Bilingualism and working memory capacity: A
 552 comprehensive meta-analysis. Second Language Research. 2017;33(3): 325-40. doi:
 553 10.1177/0267658316678286.
- 554 <u>23.35.</u> Jeon EH, Yamashita J. L2 reading comprehension and its correlates: A meta-555 analysis. Language Learning. 2014;64(1):160-212. doi: 10.1111/lang.12034.
- Linck JA, Osthus P, Koeth JT, Bunting MF. Working memory and second language
 comprehension and production: A meta-analysis. Psychonomic Bulletin & Review.
 2014;21(4): 861-83. doi: 10.3758/s13423-013-0565-2.
- 559 25.37. Bailey H, Dunlosky J, Kane MJ. Contribution of strategy use to performance on
 560 complex and simple span tasks. Memory & cognition. 2011;39(3): 447-61. doi:
 561 10.3758/s13421-010-0034-3.
- 562 <u>26.38.</u> Turner ML, Engle RW. Is working memory capacity task dependent?. Journal of 563 memory and language. 1989;28(2): 127-54. doi: 10.1016/0749-596X(89)90040-5.
- 564 <u>27.39.</u> Conway ARA, Kane , MJ , Bunting MF, Hambrick DZ, Wilhelm O, et al. (2005)
- 565 Working memory span tasks: A methodological review and user's guide. Psychonomic 566 Bulletin and Review 12(12): 769–786. doi: 10.3758/BF03196772.
- 28:40. Zhou H, Rossi S, Chen B. Effects of working memory capacity and tasks in
 processing L2 complex sentence: evidence from Chinese-English bilinguals. Frontiers in
 psychology. 2017;8: 595. doi: 10.3389/fpsyg.2017.00595
- 570 29.41. Reber PJ, Knowlton BJ, Squire LR. Dissociable properties of memory systems:
- 571 differences in the flexibility of declarative and nondeclarative knowledge. Behavioral
- 572 Neuroscience. 1996;110(5): 861. doi: 10.1037/0735-7044.110.5.861.

- 573 <u>30.42.</u> Squire LR. Memory systems of the brain: a brief history and current perspective.
 574 Neurobiology of learning and memory. 2004;82(3): 171-7. doi: 10.1016/j.nlm.2004.06.005
- 575 <u>31.43.</u> Eichenbaum H. Hippocampus: cognitive processes and neural representations that 576 underlie declarative memory. Neuron. 2004;44(1):109-20.
- 577 <u>32.44.</u> Squire LR. Memory systems of the brain: a brief history and current perspective.
 578 Neurobiology of learning and memory. 2004;82(3): 171-7. doi:
 579 10.1016/j.nlm.2004.06.005.
- 580 <u>33.45.</u> Knowlton BJ, Siegel AL, Moody TD. Procedural learning in humans. In Byrne JH,
 581 editor. Learning and memory: A comprehensive reference. 2nd ed. Oxford: Academic
 582 Press; 2017. pp. 295–312.
- 583 <u>34.46.</u> Hamrick P, Lum JA, Ullman MT. Child first language and adult second language
 584 are both tied to general-purpose learning systems. Proceedings of the National Academy
 585 of Sciences. 2018;115(7): 1487-1492. doi: 10.1073/pnas.1713975115.
- 586 <u>35.47.</u> Ullman MT. The declarative/procedural model: A neurobiologically motivated
 587 theory of first and second language. In: VanPatten B, Williams J, editors. Theories in
 588 second language acquisition: An introduction. 2nd ed. New York: Routledge; 2015. pp.
 589 135-158.
- 590 <u>36.48.</u> Ullman MT. The declarative/procedural model: A neurobiological model of
 591 language learning, knowledge, and use. In: Hickok G, Small SA, editors. Neurobiology of
 592 language. Amsterdam: Elsevier; 2016. pp. 498–505.
- 593 <u>37.49.</u> Morgan-Short K, Faretta-Stutenberg M, Brill-Schuetz KA, Carpenter H, Wong PC.
- 594 Declarative and procedural memory as individual differences in second language

- 595 acquisition. Bilingualism: Language and Cognition. 2014;17(1):56-72. doi:
 596 10.1017/S1366728912000715.
- 597 <u>38.50.</u> Carpenter, HS. A behavioral and electrophysiological investigation of different
 598 aptitudes for L2 grammar in learners equated for proficiency level. Ph.D. Thesis,
 599 Georgetown University. 2008. Available from: http://hdl.handle.net/10822/558127.
- 600 <u>39.51.</u> Carroll JB, Sapon SM. Modern Language Aptitude Test: Manual. New York:
 601 Psychological Corporation; 1959.
- 602 40.52. Trahan DE, Larrabee GJ. Continuous visual memory test. Odessa, FL: Assessment
 603 Resources. 1988.
- 604 <u>41.53.</u> Schneider W, Eschman A, Zuccolotto A. EPrime user's guide. Pittsburgh, PA:
 605 Psychology Software Tools Inc. 2002.
- 42.54. Unsworth N, Heitz RP, Schrock JC, Engle RW. An automated version of the
 operation span task. Behavior Research Methods. 2005;37(3): 498-505. doi:
 10.3758/BF03192720.
- 609 <u>55.</u> Wickens TD. Elementary signal detection theory. New York: Oxford University Press;
 610 2002.
- 611 <u>56. R Development Core Team (2016) R: A language and environment for statistical</u>
- 612 <u>computing. R Foundation for Statistical Computing, Vienna,</u>
- 613 <u>Austria. http://www.rproject.org.</u>
- 614 43.57. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using
 615 lme4. Journal of Statistical Software. 2015;67(1), 1–48. doi: 10.18637/jss.v067.i01
- 616 58. Kane MJ, Hambrick DZ, Tuholski SW, Wilhelm O, Payne TW, Engle RW. The generality
- of working memory capacity: a latent-variable approach to verbal and visuospatial memory

618 span and reasoning. Journal of Experimental Psychology: General. 2004;133(2): 189-217. 619

doi: 10.1037/0096-3445.133.2.189.

620 44.

621 45. Marsden E, Morgan Short K, Thompson S, Abugaber D. Replication in second language 622 research: Narrative and systematic reviews and recommendations for the field. Language 623 Learning. 2018;68(2): 321-391. doi: 10.1111/lang.12286.

- 624 46. Marsden E, Morgan- Short K, Thompson S, Abugaber D. Replication in second language 625 research: Narrative and systematic reviews and recommendations for the field. Language Learning. 1;68(2): 321-391. doi: 10.1111/lang.12286. 626
- 627 Gelman A. The failure of null hypothesis significance testing when studying 47.59. 628 incremental changes, and what to do about it. Personality and Social Psychology Bulletin. 629 2018;44(1): 16-23. doi: 10.1177/0146167217729162.
- 630 Leidheiser W, Branyon J, Baldwin N, Pak R, McLaughlin A. Lessons learned in 48.60. 631 adapting a lab-based measure of working memory capacity for the web. In: Proceedings of
- 632 the Human Factors and Ergonomics Society Annual Meeting 2015. Los Angeles: Sage CA;
- 633 2015. pp. 756-760.
- 634 61. Reips UD, Krantz JH. Conducting true experiments on the Web. In: Gosling SD, Johnson
- 635 JA, editors. Advanced methods for conducting online behavioral research. Washington,
- 636 DC: American Psychological Association; 2010. pp. 193-216.
- MacWhinney B. A shared platform for studying second language acquisition. 637 49.62.
- 638 Language Learning. 2017;67(S1): 254-75. doi: 10.1111/lang.12220.

- 639 <u>50.63.</u> Meurers D, Dickinson M. Evidence and interpretation in language learning
 640 research: Opportunities for collaboration with computational linguistics. Language
 641 Learning. 2017;67(S1): 66-95. doi: 10.1111/lang.12233.
- 642 <u>51.64.</u> Ziegler N, Meurers D, Rebuschat P, Ruiz S, Moreno- Vega JL, Chinkina M, Li W,
- 643 Grey S. Interdisciplinary research at the intersection of CALL, NLP, and SLA:
- 644 Methodological implications from an input enhancement project. Language Learning.
- 645 2017;67(S1): 209-231. doi: 10.1111/lang.12227.



2nd November 2019

EBERHARD KARLS

UNIVERSITÄT

TÜBINGEN

Thank you for the specific feedback on the manuscript entitled "**Measuring individual differences in cognitive abilities in the lab and on the web**". Here is our response on how we took the feedback into account in revising the paper:

Reviewer #2: Thank you for the opportunity to review this paper. It is an interesting study that compares lab-based and web-based versions of memory tests in a sample of adults with the aim of validating the web-based version.

The article is well-written and the study is set up well in general. I listed a few specific comments below:

*P3, l.59: when referring to the benefits of web-based testing it would be interesting to refer to other possible simultaneous testing strategies available. For instance, there are many tests that can be answered by individuals in school or university settings that might have similar benefits compared to web-based versions, so it would be important to emphasize what is the specific advantage of this type of tool.

While the established paper-and-pencil tests naturally can be administered by individuals in a formal education setting, conducting such tests during class time instead of conducting individualized web-based testing outside of class uses up class time that could be used for teaching and learning activities. Conducting such paper-and-pencil tests in class would also be more of an issue in school cultures in which standardized testing is less common than in the US. We added a new paragraph that discusses other methodological advantages of (remote) web-based testing in comparison to other forms of simultaneous delivery of tests, such as traditional paper-pencil and (offline) computer-based testing (p. 3).

*P4, 1.75: please provide an argument of why are you only looking at one type of equivalence.

The following argument was added (p. 5):

Considering that this study is a subcomponent of the dissertation research of the first author, limiting funding and time (see limitations below), we focused the investigation on one type of measurement equivalence, the first type: Do people who have relatively high values in one of tests also have relatively high values on the other test, and the other way around?

*P5, l.92: throughout the paper there are several mentions to L2 research, however the issue of small sample size and low power are not restricted to that research area. I would expand the claim to many other situations where methodological issues related to testing are a challenge.





The discussion of the methodological issues was expanded, including reference to low statistical power and small sample sizes being problematic in other research fields and the ongoing debate in the so-called replication crisis in psychology (p. 5-6).

*P8, l.165: the fact that the sample was not full due to technical reasons requires more explanation. Is this related to possible flaws of web-based testing? If so, it should be included in the discussion.

We added the following explanation (p. 9):

Additionally, participant numbers differed across test versions due to technical difficulties (i.e., participants erroneously entered their responses using the keyboard [Web-based CVMT]; and data was missing for one participant [Web-based MLAT5]; see description and Table 1 below, and Discussion).

and a discussion of these technical shortcomings is included in the Discussion section (p. 18).

*Dicussion: I think it would be important to discuss the limitations of the study and also of the findings.

We added limitations of the study and findings in the Discussion section (p. 18).

Yours sincerely,

Simón Ruiz, Xiaobin Chen, Patrick Rebuschat, and Detmar Meurers