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, 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, l.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
Additional Information:	
Question	**Response**
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed	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 *PLOS ONE* 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	
Competing Interests 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 *PLOS ONE* for specific examples.	The authors have declared that no competing interests exist.

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 the study involved: • Human participants • Human specimens or tissue • Vertebrate animals or cephalopods • Vertebrate embryos or tissues • Field 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.**	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.

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 PLOS Data Policy 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: *All XXX files are available from the XXX database (accession number(s) XXX, XXX.*). • If the data are all contained **within the manuscript and/or Supporting Information files**, enter the following: *All relevant data are within the manuscript and its Supporting Information files.* • If neither of these applies but you are able to provide **details of access elsewhere**, with or without limitations, please do so. For example: *Data cannot be shared publicly because of [XXX]. Data are available from the XXX Institutional Data Access / Ethics Committee (contact via XXX) for 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*	All relevant data are within the manuscript and its Supporting Information files.

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:	

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

LEAD
Graduate School &
Research Network

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

Measuring individual differences in cognitive abilities in the lab and on the web

Simón Ruiz[1*], Xiaobin Chen[2], Patrick Rebuschat[1,3] Detmar Meurers[1,4]

[1] ¶LEAD Graduate School and Research Network, University of Tübingen, Tübingen, Germany

[2] ¶Department of Theoretical and Applied Linguistics, University of Cambridge, United Kingdom

[3] ¶Department of Linguistics and English Language, Lancaster University, Lancaster, United Kingdom

[4]Department of Linguistics, University of Tübingen, Tübingen, Germany

* Corresponding author
E-mail: simon.ruiz-hernandez@sfs.uni-tuebingen.de (SR)

¶These authors contributed equally to this work.

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.

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).

The effect of working memory and declarative memory on language learning has been primarily studied in lab settings, i.e., in well-controlled environments where participants are tested individually. While this choice is methodologically sound, it can also negatively affect sample size and population sampling [13, 14]. Lab-based testing generally means testing participants individually and sequentially, which is labor-intensive and could explain why lab studies tend to have (too) few participants to allow for meaningful generalization. For example, Plonsky [13] found that the typical sample size in L2 studies was 19 participants, and Lindstromberg [15] recently reported a similar small average sample size of 20 participants. Moreover, many (if not most) lab studies in L2 research draw their sample from the surrounding student population, which is understandable given the ease of access, but also means that samples are often not representative of the population of interest.

Conducting second language research by means of remote testing via the web could alleviate some of these concerns. For example, web-based testing facilitates the acquisition of large amounts of data since participants can be tested simultaneously, and test administration can also be more cost-effective than research conducted in the lab [15]. Importantly, web-based 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 contribute directly to our understanding of individual differences in L2 acquisition. To support such task sharing and use, it is essential to first establish the validity of the online versions of the tasks (on a par with what is established about the offline versions). With this in mind, the study set out to establish measurement equivalence between lab-based and web-based tests of working memory and declarative memory.

According to Gwaltney, Shields and Shiffman ([19], 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 is related to whether differences found in one measurement are also systematically found in the other. This means that, although the two measurements estimate two different numbers, these numbers have a systematic and very clear relationship to each other. The second type concerns whether two measurements yield the same numbers. Here, we focus on the former type of equivalence. More specifically, we compare the differential performance generated by two versions of tests measuring working memory and declarative memory capacities in lab-based and web-based settings, with the aim to determine whether the two versions are equivalent with respect to the relationships between scores. Establishing measurement equivalence between these two administration modes is essential for several reasons. First, it is necessary to show that the results of web-based studies are comparable to those of previous research, which have predominantly obtained from data gathered in lab-based settings. Second, it is imperative to ensure that cognitive constructs are measured in the same way in both test modes. Finally, it is important to ascertain whether lab-based and web-based measures 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].

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 visual-spatial 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.

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

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

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, 38]. Several studies (e.g., [2, 9, 38, 39]) 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., 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]).

The present study

The main goal of the present study was to provide web-based versions of commonly employed individual difference measures in second language research, in order to make them usable in large-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.

Methods

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.

Participants

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

Materials

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

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

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

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 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.

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

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.

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

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

Descriptive statistics

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

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

Test	N	*M*	*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.

Correlations

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

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

Test	OSpan Lab-based	OSpan Web-based	MLAT5 Lab-based	MLAT5 Web-based	CVMT 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.

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

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.

Reliability

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

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.

Regression analysis

The results of the regression analyses are displayed in Table 4. For the working memory test (OSpan), the unstandardized coefficient was .89 (β = .77, *SE* = 0.10, p < .001). For the verbal declarative memory test (MLAT5), the unstandardized coefficient was .83 (β = .78, *SE* = 0.09, p < .001). And for the nonverbal declarative memory test (CVMT), the unstandardized coefficient was .74 (β = .54, *SE* = 0.19, p < .001). Overall, the results indicated that the lab-based and web-based scores are substantially related.

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

Test	Unstandardized coefficient[a]	*SE*	*p*
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. [a]The standardized coefficient (β) in parentheses.

Discussion

Studies on individual differences in language learning frequently assess the working memory and declarative memory capacities of their participants in order to determine the effect of these cognitive variables on learning outcomes. Most of this research, however, is conducted in lab-based settings, which often implies relatively small sample size and a restricted population sampling. Both of these methodological challenges can be addressed by means of remote testing via the web. In the present study, we compared lab-based and web-based individual difference measures in order to validate web-based tests for future research. The type of comparison contributes significantly to ongoing efforts to improve the methodological robustness of current second language research [47]. If web-based testing can be shown to yield comparable results to lab-based testing, researchers will be able to reach more participants for their studies, which, in turn, can help alleviate some of the current concerns in L2 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 [16, 48].

The results indicated that the scores in the lab-based and web-based versions of three cognitive tests (MLAT5, CVMT, OSpan) were equivalent in the sense that differences in performance were constant in the two versions. This suggests that participants who had relatively high values in one task also had relatively high values in the second, or the other way around. However, the strength of the association depended on the test. In both the working memory test (OSpan) and the verbal declarative memory test (MLAT5) the scores were more strongly correlated (β = .77 and β = .78, respectively); for the nonverbal declarative test (CVMT), equivalence appears to be weaker (β = .54). On the whole, the correlations reported here between lab-based and web-based scores are consistent with the assumption that both versions seem to likely measure the same cognitive construct, at least for the working memory test (OSpan) and the verbal declarative memory test (MLAT5), and, to a lesser extent, for the nonverbal 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 used left ("NEW") or right ("OLD") mouse clicking to enter their response, whereas in the web-based version, they used the keyboard ("N" and "O" keys). This modification was made to the web-based version because of technical reasons, i.e., the browser window may not register the participants' response if the cursor is not over a certain area on the page, which in itself may cause problems of missing data. It has been previously reported that participants in web-based research are prone to make errors when using the keyboard to enter their responses [49], which in this case might have affected the results of the comparison between lab-based and web-based versions of CVMT. Further studies comparing performance between the two versions may benefit from gathering data via touch input instead, which might overcome the technical difficulty of employing mouse clicking for web-based data collection reported here.

Conclusion

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 possible 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 and not comparing individual scores from one mode with scores from the other. Moreover, the extent to which the measures are equivalent varies according to the test. In this sense, we are confident that the two versions for the working memory test (OSpan) and the verbal declarative memory (MLAT5) are fairly possibly measuring the same construct, but we refrain from making such a strong statement for the nonverbal declarative test (CVMT), where the two modes might still plausibly measure strongly different aspects as well. Our research has shown that collecting experimentally controlled data on cognitive individual differences typically used in L2 research in the Internet is feasible and comparable to lab-based collection. Consequently, some of these web-based versions could very well be incorporated, for example, in future web-based intervention studies on second language learning, thereby contributing to the scaling up of data collection in the field [50-52].

Acknowledgements

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

References

1. Kidd E, Donnelly S, Christiansen MH. Individual differences in language acquisition and processing. Trends in cognitive sciences. 2018;22(2):154-69. doi: 10.1016/j.tics.2017.11.006
2. 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. 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. 396-417.
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.
6. Larsen- Freeman D. Looking ahead: Future directions in, and future research into, second 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.
8. Lado B. Aptitude and pedagogical conditions in the early development of a nonprimary language. Applied Psycholinguistics. 2017;38(3):679-701.
9. 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.
11. Buffington J, Morgan-Short K. Declarative and procedural memory as individual differences in second language aptitude. In: Wen Z, Skehan P, Biedroń A, Li S, Sparks R,

editors. Language aptitude: Multiple perspectives and emerging trends. New York: Routledge; 2019. pp. 215–237.

12. 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.
13. Plonsky L. Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. Studies in Second Language Acquisition. 2013;35(4): 655-87. doi: 10.1017/S0272263113000399
14. Plonsky L. Quantitative research methods. In: Loewen S, Sato M, editors. The Routledge handbook of instructed second language acquisition. New York: Routledge; 2017. pp. 505-521.
15. Lindstromberg S. Inferential statistics in Language Teaching Research: A review and ways forward. Language Teaching Research. 2016;20(6): 741-68. doi: 10.1177/1362168816649979.
16. Krantz JH, Reips UD. The state of web-based research: A survey and call for inclusion in curricula. Behavior Research Methods. 2017;49(5): 1621-1619. doi: 10.3758/s13428-017-0882-x
17. Hicks KL, Foster JL, Engle RW. Measuring working memory capacity on the web with the online working memory lab (the OWL). Journal of Applied Research in Memory and Cognition. 2016;5(4): 478-89. doi: 10.1016/j.jarmac.2016.07.010.
18. Wolfe CR. Twenty years of Internet-based research at SCiP: A discussion of surviving concepts and new methodologies. Behavior research methods. 2017;49(5): 1615-1620. doi: 10.3758/s13428-017-0858-x.

19. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. Value in Health. 2008;11(2): 322–333. doi: 10.1111/j.1524-4733.2007.00231.x.

20. Cowan N. Working memory maturation: Can we get at the essence of cognitive growth?. Perspectives on Psychological Science. 2016;11(2): 239-64. doi: 10.1177/1745691615621279

21. Baddeley AD. Modularity, working memory and language acquisition. Second Language Research. 2017;33(3): 299-311. doi: 10.1177/0267658317709852.

22. Roehr K. Linguistic and metalinguistic categories in second language learning. Cognitive Linguistics. 2008;19(1): 67-106. doi: 10.1515/COG.2008.005.

23. Grundy JG, Timmer K. Bilingualism and working memory capacity: A comprehensive meta-analysis. Second Language Research. 2017;33(3): 325-40. doi: 10.1177/0267658316678286.

24. Jeon EH, Yamashita J. L2 reading comprehension and its correlates: A meta- analysis. Language Learning. 2014;64(1):160-212. doi: 10.1111/lang.12034.

25. 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.

26. 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/s13421-010-0034-3.

27. 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.

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.

29. 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

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.

31. 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

32. Eichenbaum H. Hippocampus: cognitive processes and neural representations that underlie declarative memory. Neuron. 2004;44(1):109-20.

33. 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.

34. 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.

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.

37. 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.

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.

39. 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.

40. Carroll JB, Sapon SM. Modern Language Aptitude Test: Manual. New York: Psychological Corporation; 1959.

41. Trahan DE, Larrabee GJ. Continuous visual memory test. Odessa, FL: Assessment Resources. 1988.

42. Schneider W, Eschman A, Zuccolotto A. EPrime user's guide. Pittsburgh, PA: Psychology 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.

44. Wickens TD. Elementary signal detection theory. New York: Oxford University Press; 2002.

45. 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. 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.

47. 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. 1;68(2): 321-391. doi: 10.1111/lang.12286.

48. 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.

49. Leidheiser W, Branyon J, Baldwin N, Pak R, McLaughlin A. Lessons learned in adapting a lab-based measure of working memory capacity for the web. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2015. Los Angeles: Sage CA; 2015. pp. 756-760.

50. MacWhinney B. A shared platform for studying second language acquisition. Language Learning. 2017;67(S1): 254-75. doi: 10.1111/lang.12220.

51. Meurers D, Dickinson M. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. Language Learning. 2017;67(S1): 66-95. doi: 10.1111/lang.12233.

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

Measuring individual differences in cognitive abilities in the lab and on the web

Simón Ruiz[1*], Xiaobin Chen[1], Patrick Rebuschat[1,2] Detmar Meurers[1,3]

[1 ¶]LEAD Graduate School and Research Network, University of Tübingen, Tübingen, Germany

[2 ¶]Department of Linguistics and English Language, Lancaster University, Lancaster, United Kingdom

[3]Department of Linguistics, University of Tübingen, Tübingen, Germany

* Corresponding author
E-mail: simon.ruiz-hernandez@sfs.uni-tuebingen.de (SR)

[¶]These authors contributed equally to this work.

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.

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).

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

The present study compared lab-based and web-based versions of cognitive tests that are widely used in disciplines such as psychology and second language research. Particularly, our 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 research community, and especially to researchers working on the area of L2 acquisition. 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 alleviating the data collection issues associated with lab-based research . Moreover, making these specific tasks available will also contribute directly to our understanding of individual differences in L2 acquisition. To support such task sharing and use, it is essential to first establish the validity of the online versions of the tasks (on a par with what is established about the offline versions). With this in mind, the study set out to establish measurement equivalence 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 associated with each other. The second type concerns whether two measurements yield the same numbers. 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? More specifically, we compare the differential performance generated by two versions of tests measuring working memory and declarative memory abilities in lab-based and web-based settings, in order to assess whether the two versions are equivalent regarding the relationships between scores.

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

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.

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

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

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]).

The present study

The main goal of the present study was to provide web-based versions of commonly employed individual difference measures in second language research, in order to make them usable in large-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.

Methods

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.

Participants

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

Materials

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

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

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

Nonverbal declarative memory. The Continuous Visual Memory Task (CVMT; [52]) served as a measure of nonverbal declarative memory [9, 49, 50]. As a visual recognition test, the CVMT is entails asking participants to first view a collection of complex abstract designs on the screen, and then to indicate whether the image they just saw was novel ("new") in the collection, or they had seen the image before ("old"). Seven of the designs were "old" (target items), and 63 were "new" (distractors). The target items appeared seven times (49 trials), and the distractors only once (63 trials) across the test. All items were shown in a random but fixed order, each one appearing on the screen for two seconds. Following the two seconds, participants were instructed to respond to the "OLD or NEW?" prompt on the screen. In the lab-based mode, 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

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

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.

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

Results

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

Descriptive statistics

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

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

Test	N	*M*	*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.

Correlations

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

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

Test	OSpan Lab-based	OSpan Web-based	MLAT5 Lab-based	MLAT5 Web-based	CVMT 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.

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

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.

Reliability

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

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.

Regression analysis

The results of the regression analyses are displayed in Table 4. For the working memory test (OSpan), the unstandardized coefficient was .89 (β = .77, SE = 0.10, p < .001). For the verbal declarative memory test (MLAT5), the unstandardized coefficient was .83 (β = .78, SE = 0.09, p < .001). And for the nonverbal declarative memory test (CVMT), the unstandardized coefficient was .74 (β = .54, SE = 0.19, p < .001). Overall, the results indicated that the lab-based and web-based scores are substantially related.

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

Test	Unstandardized coefficient[a]	*SE*	*p*
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. [a]The standardized coefficient (β) in parentheses.

Discussion

Studies on individual differences in language learning frequently assess the working memory and declarative memory capacities of their participants in order to determine the effect of these cognitive variables on learning outcomes. Most of this research, however, is conducted in lab-based settings, which often implies relatively small sample size and a restricted population sampling. Both of these methodological challenges can be addressed by means of remote testing via the web. In the present study, we compared lab-based and web-based individual difference measures in order to validate web-based tests for future research. The type of comparison contributes significantly to ongoing efforts to improve the methodological robustness of current second language research, for example [12]. If web-based testing can be shown to yield comparable results to lab-based testing, researchers will be able to reach more participants for their studies, which, in turn, can help alleviate some of the current concerns in lab-based 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 [17, 59].

Findings showed that the scores in the lab-based and web-based versions of three cognitive tests (MLAT5, CVMT, OSpan) were equivalent concerning differences in performance, which were constant in the two versions, suggesting that participants who had relatively high values in one task also had relatively high values in the second, or the other way around. However, the strength of the relationship was a function of the kind of test. More specifically, in both the working memory test (OSpan) and the verbal declarative memory test (MLAT5), the scores were more strongly correlated (β = .77 and β = .78, respectively); for the nonverbal declarative test (CVMT), equivalence appears to be weaker (β = .54).Overall, the correlations reported here between lab-based and web-based scores are consistent with the assumption that both versions seem to likely measure the same cognitive construct, at least for the working memory test (OSpan) and the verbal declarative memory test (MLAT5), and, to a 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 the web-based version was motivated by technical reasons, specifically, the browser window may not register the participants’ response if the cursor is not over a certain area on the page, which in turn may cause problems of missing data. Previous research has found that participants in web-based research are particularly prone to err when using the keyboard to input their responses [60], which in this case might have affected the results of the comparison between lab-based and web-based versions of CVMT. Future research comparing performance between the lab and web-versions may benefit from collecting data through touch input instead, as this might help overcome potential technical difficulties caused by using mouse clicking for web-based data.

Some limitations of the study and the findings presented here should be considered. One of the limitations was the small sample size. As mentioned earlier, logistic constrains due to the availability of time and funding prevented the researchers from testing more participants for this study. In addition, the fact that some participants (14%) dropped out before completing any of the web-based measures in the second part of the experiment, which is typical in web-based research [17], also contributed to the reduction of the data available for the comparison between lab and web-based testing in the present investigation. Therefore, our findings should be replicated in a larger study. A second limitation was that test-retest reliability was not examined here, given that the main aim of this study was to establish valid online versions of known individual difference measures. Future research should assess test-retest reliability, as it is as an interesting endeavor for studying individual difference measures in future work. Finally, and as indicated above, a third limitation concerned technical issues that affected data collection, as some participants used the wrong keys on the keyboard to submit their responses to the web-based version of the CVMT, rendering the data from some of the participants impossible to use for the comparison; furthermore, data from one subject was missing in the Web-based MLAT, which may have been due to technical issues at the participant's end (e.g., not following the general instructions given, such as refreshing or closing the browser page [see Procedure]; or Internet disconnection). In this sense, Reips and Krantz [61] (see also[17]) caution researchers that one of the potential disadvantages of Internet-driven testing is the technical variability characteristic of web-based research (e.g., different browsers and Internet connections), which, in turn, may affect data collection.

Conclusion

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 possible 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 and not comparing individual scores from one mode with scores from the other. Moreover, the extent to which the measures are equivalent varies according to the test. In this sense, we are confident that the two versions for the working memory test (OSpan) and the verbal declarative memory (MLAT5) are likely to measure the same construct, whereas the correlation between the nonverbal declarative test (CVMT) versions was less pronounced. Our research has shown that collecting experimentally controlled data on cognitive individual differences typically used in the area of L2 research in the Internet is feasible and comparable to lab-based collection. Consequently, some of these web-based versions could very well be incorporated, for example, in future web-based intervention studies on second language learning, thereby contributing to the scaling up of data collection in the field [62-64].

Acknowledgements

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

References

1. Kidd E, Donnelly S, Christiansen MH. Individual differences in language acquisition and processing. Trends in Cognitive Sciences. 2018;22(2):154-69. doi: 10.1016/j.tics.2017.11.006
2. 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. 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. 396-417.
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.

6. Larsen- Freeman D. Looking ahead: Future directions in, and future research into, second 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.
8. Lado B. Aptitude and pedagogical conditions in the early development of a nonprimary language. Applied Psycholinguistics. 2017;38(3):679-701.
9. 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.
11. Buffington J, Morgan-Short K. Declarative and procedural memory as individual differences in second language aptitude. In: Wen Z, Skehan P, Biedroń A, Li S, Sparks R, editors. Language aptitude: Multiple perspectives and emerging trends. New York: Routledge; 2019. pp. 215–237.
12. 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.

13. Plonsky L. Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. Studies in Second Language Acquisition. 2013;35(4): 655-87. doi: 10.1017/S0272263113000399

14. Plonsky L. Quantitative research methods. In: Loewen S, Sato M, editors. The Routledge handbook of instructed second language acquisition. New York: Routledge; 2017. pp. 505-521.

15. Lindstromberg S. Inferential statistics in Language Teaching Research: A review and ways forward. Language Teaching Research. 2016;20(6): 741-68. doi: 10.1177/1362168816649979.

16. Tackett JL, Brandes CM, King KM, Markon KE. Psychology's replication crisis and clinical psychological science. Annual review of clinical psychology. 2019;15:579-604.

17. Krantz JH, Reips UD. The state of web-based research: A survey and call for inclusion in curricula. Behavior Research Methods. 2017;49(5): 1621-1619. doi: 10.3758/s13428-017-0882-x

18. Roever C. Web-based language testing. Language Learning & Technology. 2001;5(2):84-94.

19. Domínguez C, López-Cuadrado J, Armendariz A, Jaime A, Heras J, Pérez TA. Exploring the differences between low-stakes proctored and unproctored language testing using an Internet-based application. Computer Assisted Language Learning. 2019:1-27.

20. Diaz Maggioli GH. Web- Based Testing. The TESOL Encyclopedia of English Language Teaching. 2018:1-6.

21. Birnbaum MH. Human research and data collection via the Internet. Annu. Rev. Psychol.. 2004;55:803-32.

22. Hicks KL, Foster JL, Engle RW. Measuring working memory capacity on the web with the online working memory lab (the OWL). Journal of Applied Research in Memory and Cognition. 2016;5(4): 478-89. doi: 10.1016/j.jarmac.2016.07.010.

23. Wolfe CR. Twenty years of Internet-based research at SCiP: A discussion of surviving concepts and new methodologies. Behavior research methods. 2017;49(5): 1615-1620. doi: 10.3758/s13428-017-0858-x.

24. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. Value in Health. 2008;11(2): 322–333. doi: 10.1111/j.1524-4733.2007.00231.x.

25. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience. 2013;14(5):365.

26. Branch MN. The "Reproducibility Crisis:" Might the Methods Used Frequently in Behavior-Analysis Research Help?. Perspectives on Behavior Science. 2019;42(1):77-89.

27. Laraway S, Snycerski S, Pradhan S, Huitema BE. An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis. Perspectives on Behavior Science. 2019;42(1):33-57.

28. Shrout PE, Rodgers JL. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. Annual review of psychology. 2018;69:487-510.

29. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 2013.

30. Stewart N, Chandler J, Paolacci G. Crowdsourcing samples in cognitive science. Trends in cognitive sciences. 2017;21(10):736-48

31. Cowan N. Working memory maturation: Can we get at the essence of cognitive growth?. Perspectives on Psychological Science. 2016;11(2): 239-64. doi: 10.1177/1745691615621279

32. Baddeley AD. Modularity, working memory and language acquisition. Second Language 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.

34. Grundy JG, Timmer K. Bilingualism and working memory capacity: A comprehensive meta-analysis. Second Language Research. 2017;33(3): 325-40. doi: 10.1177/0267658316678286.

35. Jeon EH, Yamashita J. L2 reading comprehension and its correlates: A meta- analysis. Language Learning. 2014;64(1):160-212. doi: 10.1111/lang.12034.

36. 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.

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/s13421-010-0034-3.

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.

39. 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.

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

41. 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.

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

43. Eichenbaum H. Hippocampus: cognitive processes and neural representations that underlie 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. 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.

60. Leidheiser W, Branyon J, Baldwin N, Pak R, McLaughlin A. Lessons learned in adapting a lab-based measure of working memory capacity for the web. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2015. Los Angeles: Sage CA; 2015. pp. 756-760.

61. Reips UD, Krantz JH. Conducting true experiments on the Web. In: Gosling SD, Johnson JA, editors. Advanced methods for conducting online behavioral research. Washington, DC: American Psychological Association; 2010. pp. 193-216.

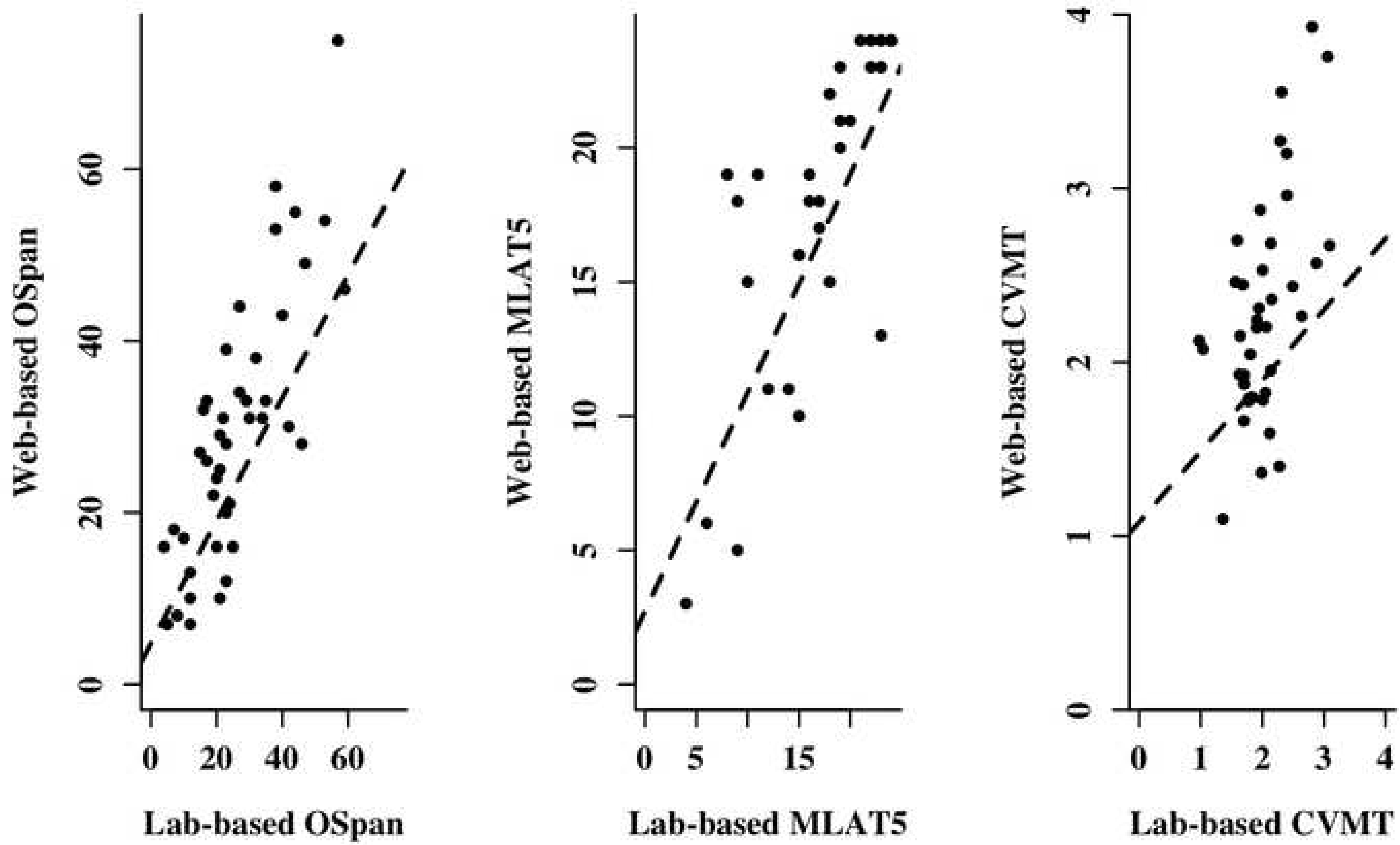
62. MacWhinney B. A shared platform for studying second language acquisition. Language Learning. 2017;67(S1): 254-75. doi: 10.1111/lang.12220.

63. Meurers D, Dickinson M. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. Language Learning. 2017;67(S1): 66-95. doi: 10.1111/lang.12233.

64. Ziegler N, Meurers D, Rebuschat P, Ruiz S, Moreno- Vega JL, Chinkina M, Li W, Grey S. Interdisciplinary research at the intersection of CALL, NLP, and SLA: Methodological

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

Fig1


Web-based OSpan
0
20
40
60
0
20
40
60
Lab-based OSpan
Web-based MLAT5
0
5
10
15
20
0
5
15
Lab-based MLAT5
Web-based CVMT
0
1
2
3
4
0
1
2
3
4
Lab-based CVMT

Measuring individual differences in cognitive abilities in the lab and on the web

Simón Ruiz[1*], Xiaobin ~~Chen[2]~~Chen[1], Patrick Rebuschat[1,~~3~~,2] Detmar Meurers[1,~~4~~3]

[1] [¶]LEAD Graduate School and Research Network, University of Tübingen, Tübingen, Germany

~~[2] [¶]Department of Theoretical and Applied Linguistics, University of Cambridge, United Kingdom~~

[~~3~~2] [¶]Department of Linguistics and English Language, Lancaster University, Lancaster, United Kingdom

~~[4]Department~~ [3]Department of Linguistics, University of Tübingen, Tübingen, Germany

* Corresponding author
E-mail: simon.ruiz-hernandez@sfs.uni-tuebingen.de (SR)

[¶]These authors contributed equally to this work.

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.

Introduction

Individual differences can greatly affect how we acquire and process language [1-3] and mediate and/or 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).

The effect of working memory and declarative memory on language learning has been primarily studied in lab settings, i.e., in well-controlled environments where participants are tested individually. While this choice is methodologically sound, it can also negatively affect sample size and population sampling [12, 13, 14~~3, 14~~]. Lab-based testing generally means testing participants individually and sequentially, which is labor-intensive and could explain why lab studies tend to have (too) few participants to allow for meaningful generalization. ~~For example~~As ~~a way of~~an example, in second language (L2) research, Plonsky [13] found that the typical sample size in L2 studies was 19 participants, and Lindstromberg [15] recently reported a similar small average sample size of 20 participants. In the same vein, [16] reported that, in psychology, median sample 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 most) lab studies in ~~L2~~ research draw their sample from the surrounding student population, which is understandable given the ease of access, but also means that samples are often not representative of the population of interest. Conducting research by means of remote testing via the web could alleviate some of these concerns. For example, web-based testing facilitates the acquisition of large amounts of data since participants can be tested simultaneously, ~~which in turn~~ enabling~~es~~ researchers to run higher~~er~~-powered studies. Likewise, test administration can also be more cost-effective than research conducted in the lab [~~15~~17]. ~~Web-based experimenting has been found to be a reliable and effective research tool [16,17, 18].~~

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 ~~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].

The present study compared lab-based and web-based versions of cognitive tests that are widely used in disciplines such as psychology and second language research. Particularly, our ~~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, and especially to researchers working on the area of L2 acquisition. T~~. T~~he 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~~ alleviating the data collection issues ~~problem~~associated with lab-based research ~~in SLA~~. Moreover, making these specific tasks available will also contribute directly to our understanding of individual differences in L2 acquisition. To support such task sharing and use, it is essential to first establish the validity of the online versions of the tasks (on a par with what is established about the offline versions). With this in mind, the study set out to establish measurement equivalence between lab-based and web-based tests of working memory and declarative memory.

According to Gwaltney, Shields and Shiffman ([~~19~~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 ~~is~~ re~~re~~gards~~lated~~ to whether differences observed ~~found~~ in one measurement are also systematically found ~~found~~ in the other, meaning that, ~~. This means that,~~ even when ~~although~~ the two measurements produce ~~estimate~~ two different numbers, these~~these~~ numbers are clearly and ~~have a~~ systematically ~~and very clear relationship~~ associated with~~to~~ each other. The second type concerns whether two measurements yield the same numbers. Considering that this study ~~was a piecework~~is a subcomponent of the dissertation research of the first author,~~ with~~ limiting~~ed~~ funding and time (see limitations below), ~~it was therefore decided to undertake a more~~we focused the investigation on ~~by looking at only~~ one type of measurement equivalence, ~~in this case,~~ 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? More specifically, we compare the differential performance generated by two versions of tests measuring working memory and declarative memory abilities ~~capacities~~ in lab-based and web-based settings, in order ~~with the aim~~ to ~~determin~~assesse whether the two versions are equivalent regarding ~~with respect to~~ the relationships between scores.

Assessing ~~measurement~~ equivalence between ~~these two administration modes~~lab and web-based measurements is essential for several reasons. Firstly, it is necessary to demonstrate ~~show~~ that the findings~~results~~ obtained in ~~of~~ web-based studies are comparable to those of previous research, which have been mainly collected ~~e predominantly obtained from data gathered~~ in lab-based settings. Secondly, it is ~~imperative~~ important to ensure ~~to ensure~~ that cognitive constructs ~~constructs~~ are similarly gauged ~~measured~~ in ~~the same way in~~ both testing modalities~~ modes~~. ~~Finally~~Likewise, 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 feasible alternative~~ to be a viable way to tackle~~address~~ some of the current methodological issues found in ~~L2~~ research conducted in lab-based settings, such as underpowered studies, restricted population sampling, and small sample sizes ~~, among others~~ [~~13~~17, ~~14~~22, 23]. Of these methodological issues, in particular, low statistical power and small sample sizes have been identified as key factors in the ongoing discussions about the reproducibility of research findings in life and social sciences [25-27]. In psychology, for example, there is currently considerable debate about the so-called *replication crisis* [28], that is, failure to reproduce significant findings when replicating previous research [27]. In this regard, and considering that much research is underpowered [29, 30], web-based testing can enable the collection of larger sample sizes, and 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].

Working memory

Working memory is ~~refers to~~ the capacity to ~~simultaneously~~ process and hold ~~retain~~ information at the same time while performing ~~carrying out~~ complex cognitive tasks such as language learning, comprehension and production [~~20~~31]. According to ~~Following~~ Baddeley and colleagues (e.g., [~~21~~32]), working memory is a multicomponent system that includes ~~consists of~~ storage subsystems ~~that are~~ responsible for retaining both ~~holding~~ visual-spatial and auditory information, an episodic buffer that serves ~~acts~~ as a link between the storage subsystems and long-term memory, and a central executive that acts~~functions~~ as an attentional control system.

Regarding ~~In~~ L2 learning, working memory ~~appears to~~ assists learners to simultaneously ~~jointly~~ process form, meaning and use of language forms ~~at the same time~~. More specifically, working memory is involved in key cognitive processes such as decision making, attention control, explicit deduction, information retrieval and analogical reasoning [4]. Moreover, working memory is also important for retaining metalinguistic information while comprehending and producing L2 language [~~22~~33]. In this regard, meta-analytic work has reported the important role of working memory in L2 comprehension and production (e.g., [~~23~~34-~~25~~36]). For example, Linck et al. ([~~25~~36], p. 873) found that working memory has a positive impact on L2 comprehension outcomes (*r* = 0.24). Likewise, Jeon and Yamashita's [~~24~~35] meta-analysis also showed that working memory is related to L2 reading comprehension (*r* = 0.42). Regarding production, meta-analytic research has, too, indicated a significant association with working memory (e.g., [~~25~~36]). In this case, Linck et al. ([~~25~~36], p. 873) found a positive correlation for productive outcomes as well (*r* = 0.27).

Working memory is often measured by means of simple or complex span tasks. Simple span tasks, such as ~~(e.g.,~~ digit span and letter span, ~~)~~ entails ~~involve~~ recalling short lists of items, and they seek to ~~gauge~~ measure the storage component ~~aspect~~ of working memory [~~26~~37]. Complex span tasks, such as the operation span task (OSpan; [~~27~~38]), on the other hand, ~~entail~~ include remembering stimuli while performing a another ~~secondary~~ task. This type of tasks ~~, and are thought to tax~~ taxes both processing (attention) and storage (memory) aspects of ~~components of~~ working memory [~~21~~32]. Here, we focus on a complex task, namely the OSpan. This complex task has been found to be a valid and reliable measure of working memory capacity [~~28~~39], and has also been recommended as a more accurate measure to examine the association between working memory and L2 processing and learning [~~29~~40].

Declarative memory

Declarative memory is the capacity to consciously recall and use information [~~30~~41]. The declarative memory system is one of the long-term memory systems in the brain [~~31~~42]. It is mainly responsible for the processing, storage, and retrieval of information about facts (semantic knowledge) and events (episodic knowledge; [~~32~~43, 4~~34~~3]). Learning in the declarative memory system is quick, intentional, and attention-driven [45~~34~~].

Substantial research has now investigated the role of declarative memory in first and second language acquisition [~~35~~46]. In first language acquisition, declarative memory is ~~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~~47, ~~37~~48]). In the case of L2 acquisition, declarative memory ~~appears to~~ underpins the learning, storage and processing of L2 vocabulary and grammar [47, 48~~36,37~~], at least in the earliest phases of acquisition [~~35~~46, ~~38~~49]. Several studies (e.g., [2, 9, ~~38~~49, 50~~39~~]) 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., ~~38~~49, 50~~39~~), both verbal, such as the paired associates subtest of the Modern Language Aptitude Test (MLAT5; [~~40~~51]), and nonverbal, such as the Continuous Visual Memory Task (CVMT; [~~41~~52]).

The present study

The main goal of the present study was to provide web-based versions of commonly employed individual difference measures in second language research, in order to make them usable in large-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.

Methods

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.

Participants

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

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

Materials

Three cognitive tests were administered, one ~~assessing~~ measuring working memory capacity, and two ~~indexing~~ assessing verbal and nonverbal declarative memory ~~capacity~~ abilities, respectively. In the lab-based ~~contex~~setting~~t~~, both working memory and nonverbal declarative memory tests were programmed and delivered via E-Prime v2.0 [~~42~~53]; the verbal declarative memory test was given~~applied~~ in paper-pencil form, as originally developed and delivered. Moreover, ~~For the w~~web-based ~~mode,~~ versions of the three cognitive tests were developed for this study using Java with the GoogleWeb Toolkit (http://www.gwtproject.org), and were accessible from all browsers. A description of each ~~The~~ test~~s~~ is given below.~~are described below.~~

Working memory. ~~To assess participants' working memory capacity, a~~An adapted version of the Automated Operation Span Task (OSpan; [~~43~~54]), a computerized form of the complex span task created by Turner and Engle [~~27~~38], was used to gauge participants' working memory capacity [9, 22~~17~~]. ~~This adaptation was b~~Based on the Klingon Span Task implemented ~~developed~~ by Hicks et al. [~~17~~22], this version ~~and~~ consisted of using Klingon symbols ~~replacing~~ instead of letters, ~~(~~the ~~original~~ stimuli to be remembered in the original OSpan task~~) with Klingon symbols~~. In Hicks et al.' study, participants cheated by writing down the letter memoranda in the web-based version of the classic OSpan, ~~causing~~motivating ~~Hicks et al. implemented t~~the~~his~~ 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.~~

The task ~~took approximately 25 minutes to complete, and~~ included ~~was divided into~~ a practice phase and a testing phase. In the practice phase, participants were first ~~presented~~ shown with a series of Klingon symbols on the screen, and then were asked to recall~~emember~~ them in the order in which they had appeared after ~~at the end of~~ each trial (i.e., symbol recall). Next, participants were required~~asked~~ to solve a series of simple ~~math operations~~equations (e.g., 8~~5~~ * 4 ~~2~~+ 7~~1~~ = ?). Finally, subjects performed the symbol recall while also solving the math problems, as they would ~~do~~ later do in the actual testing phase. Following ~~After~~ the practice phase, participants were ~~presented~~ shown with the real trials, which consisted of a list of 15 sets of 3–7 randomized symbols that appeared intermingled~~mixed~~ with the equations. In sum, there were~~, totaling~~ 75 symbols and 75 math problems. At the end of each set, participants were asked to remember~~call~~ the symbols in the sequence they had been presented~~shown~~. An individual time limit to answer the math problems in the real trials was calculated ~~derived~~ from the average response time plus 2.5 standard deviations taken during the math practice section. Following Unsworth et al. [~~46~~54], a partial score (i.e., total number of correct symbols recalled in the correct order) was taken as the OSpan score (see [~~28~~39], for a description of scoring procedures). The highest possible score was 75. The entire task took about 25 min.

Verbal declarative memory. To measure verbal declarative memory, t~~T~~he Modern Language Aptitude Test, Part 5, Paired Associates (MLAT5; [~~40~~51]), was used ~~as a verbal measure of declarative memory~~ [9, ~~38~~49, 50~~39~~]. In t~~T~~he MLAT5, ~~required~~ participants were required to memorize artificial, pseudo-Kurdish words and their meanings in English. Participants were first asked to study~~ied~~ 24-word association pairs for two minutes, and then complete~~ed~~ a two-minute practice section.~~ During the practice section,~~ T~~t~~he list of foreign words with ~~and~~ their respective English ~~equivalents~~ meanings was ~~were~~ made available for participants as they completed the practice session~~to refer back if they needed to~~. Finally, subjects were instructed to complete~~d~~ a timed multiple-choice test (four minutes), by ~~in which they were asked to~~ selecting the English meaning of each of the 24 pseudo-Kurdish words from five options previously ~~seen~~ displayed at the memorization stage. For each correct response, one point was ~~awarded~~given, yielding a total score of 24 points. The test duration was about 8 minutes.

Nonverbal declarative memory. The Continuous Visual Memory Task (CVMT; [~~46~~52]) served ~~was included~~ as a measure~~n assessment~~ of nonverbal declarative memory [9, ~~38~~49, 50 ~~39~~]. As a visual recognition test, t~~T~~he CVMT is ~~a visual recognition test that involves~~entails asking participants to first view a collection of complex abstract designs on the screen, and then to indicate whether the image they just saw was novel (“new”) in the collection, or they had seen the image before (“old”). Seven of the designs were “old” (target items), and 63 were “new” (distractors). ~~Throughout the task, t~~The target items appeared seven times (49 trials), and the distractors only once (63 trials) across the test. All items were ~~presented~~ shown in a random but fixed order, each one appearing on the screen for two seconds. ~~After~~ Following the two seconds, participants were instructed to respond to the “OLD or NEW?” prompt on the screen. In the lab-based ~~setting~~mode, subjects used mouse click for making ~~indicated~~ their choice, ~~by mouse clicking either~~ left for “NEW”, or right for “OLD”. In the web-based~~-~~ mode~~setting~~, they responded by pressing either the “N” key for “NEW”, or the “O” key for “OLD” on the keyboard. T~~Overall, t~~he CVMT took ~~required~~ 10 min~~utes~~ to ~~be~~ complete~~d~~. ~~For each participant, a~~A *d*'(d-prime) score [~~44~~55] ~~for CVMT~~ was calculated for each participant~~omputed~~. The d' score was used to ~~account for~~reduce potential ~~the possible participants'~~ response bias ~~toward choosing "OLD" or "NEW"~~.

Procedure

As previously noted, ~~As previously noted, p~~participants ~~completed~~ 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~~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. Upon finishing the MLAT5, subjects then ~~They were then asked to~~ filled out ~~in~~ a background questionnaire. The whole lab-based session lasted ~~took~~ about 40 min~~nutes~~.

~~For~~ Regarding the web-based session, each subject was sent an email with ~~containing~~ a unique web link with a personalized code, ~~that when~~which once clicked, took them to an interface that hosted ~~housing~~ the web-based versions of the cognitive tests. In order t~~T~~o avoid ~~prevent participants from taking the tests~~ multiple~~-~~ responses by the same participant~~times~~, the link was disabled ~~became nonfunctional~~ once subjects ~~they~~ had submitted their responses in the last test (i.e., MLAT5). In the email, participants were also informed that the web-based session lasted about 40 min~~utes~~, and that it had to be completed within a week. On the interface, following informed consent, subjects were ~~given~~ provided with general instructions that reflected ~~in 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 the~~at 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, t~~T~~he mean period between the first and second testing was 45.7 days (*SD* = 4.1).

Results

All data were analyzed ~~using~~ by means of ~~the statistical software package~~ R (version 3.3.2; ~~(~~[56]~~R Core Team, 2016~~). Missing data was ignored (complete-~~-~~case analysis). Linear regression models were built using the lm function in the lme4 library [57]. From a temporal perspective~~point of view~~, lab scores were used to predict web scores in the linear regression models. To verify normality, model residuals were visually inspected. Reliability was assessed using Cronbach's alpha. Following Kane et al. [~~45~~58], for the lab-based working memory test (OSpan-Lab-based), reliability was assessed by calculating the proportion of correctly recalled Klingon symbols per each of the 15 trials in the test (e.g., one out of four symbols correctly recalled corresponded to a proportion of .25). For the web-based working memory test (OSpan-Web-based), however, internal consistency is not reported, since it was not technically possible to perform a detailed item-based analysis. Descriptive statistics are presented first, followed by correlations, internal consistency estimates (Cronbach's alpha), and the results of linear regression analyses.

Descriptive statistics

Table 1 presents the descriptive statistics for ~~summarizing participants’~~ participants’ performance on ~~the three~~ cognitive tests ~~under investigation~~ in both testing- settings~~modes~~.

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

Test	N	*M*	*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.

Correlations

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

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

Test	OSpan Lab-based	OSpan Web-based	MLAT5 Lab-based	MLAT5 Web-based	CVMT 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.

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

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.

Reliability

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

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.

Regression analysis

The results of the regression analyses are displayed in Table 4. For the working memory test (OSpan), the unstandardized coefficient was .89 (β = .77, *SE* = 0.10, $p < .001$). For the verbal declarative memory test (MLAT5), the unstandardized coefficient was .83 (β = .78, *SE* = 0.09, $p < .001$). And for the nonverbal declarative memory test (CVMT), the unstandardized coefficient was .74 (β = .54, *SE* = 0.19, $p < .001$). Overall, the results indicated that the lab-based and web-based scores are substantially related.

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

Test	Unstandardized coefficient[a]	*SE*	*p*
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. [a]The standardized coefficient (β) in parentheses.

Discussion

Studies on individual differences in language learning frequently assess the working memory and declarative memory capacities of their participants in order to determine the effect of these cognitive variables on learning outcomes. Most of this research, however, is conducted in lab-based settings, which often implies relatively small sample size and a restricted population sampling. Both of these methodological challenges can be addressed by means of remote testing via the web. In the present study, we compared lab-based and web-based individual difference measures in order to validate web-based tests for future research. The type of comparison contributes significantly to ongoing efforts to improve the methodological robustness of current second language research, for example [~~47~~12]. If web-based testing can be shown to yield comparable results to lab-based testing, researchers will be able to reach more participants for their studies, which, in turn, can help alleviate some of the current concerns in ~~L2~~ lab-based 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 [~~16~~17, ~~48~~59].

Findings ~~The results~~ showed~~indicated~~ that the scores in the lab-based and web-based versions of three cognitive tests (MLAT5, CVMT, OSpan) were equivalent ~~in the sense that~~concerning differences in performance, which were constant in the two versions, ~~. This~~ suggestings~~s~~ that participants who had relatively high values in one task also had relatively high values in the second, or the other way around. However, the strength of the ~~association~~ relationship was a function of ~~depended on~~ the kind of test. More specifically, i~~I~~n both the working memory test (OSpan) and the verbal declarative memory test (MLAT5), the scores were more strongly correlated (β = .77 and β = .78, respectively); for the nonverbal declarative test (CVMT), equivalence appears to be weaker (β = .54). Overall~~On the whole~~, the correlations reported here between lab-based and web-based scores are consistent with the assumption that both versions seem to likely measure the same cognitive construct, at least for the working memory test (OSpan) and the verbal declarative memory test (MLAT5), and, to a lesser extent, for the nonverbal declarative test (CVMT).

A ~~possible~~ potential explanation for lesser ~~the weaker~~ equivalence in ~~found for~~ the versions~~s~~ of the nonverbal declarative test (CVMT) ~~is perhaps~~could be due to the different ~~difference in the way~~manner in which ~~r~~the 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 left (“NEW”) or right (“OLD”) mouse clicking to provide a~~enter~~ ~~their~~ response, whereas in the web-based version, they used the keyboard (“N” and “O” keys). This ~~is~~ modification made to ~~was made to~~ the web-based version was motivated by ~~because of~~ technical reasons, specifically, ~~, i.e.,~~ the browser window may not register the participants’ response if the cursor is not over a certain area on the page,~~ which in itself~~ which in turn may ~~may~~ cause problems of missing data. Previous research ~~It has been previously reported~~has found that participants in web-based research are particularly prone ~~prone~~ to err ~~make errors~~ when using the keyboard to input ~~enter~~ their responses [~~49~~60], which in this case might have affected the results of the comparison between lab-based and web-based versions of CVMT. Future research ~~Further studies~~ comparing performance between the ~~two versions~~lab and web-versions may benefit from collecting ~~gathering~~ data ~~vi~~through~~a~~ touch input instead, as this~~which~~ might help overcome potential ~~the~~ technical difficulties~~y~~ ~~of~~ caused by using ~~employing~~ mouse clicking for web-based data. ~~collection reported here.~~

~~Limitations~~

Some limitations of the study and the findings presented here should be considered. One of the limitations was the small sample size. As mentioned earlier, logistic constrains due to the availability of time and funding prevented the researchers from testing more participants for this study. In addition, the fact that some participants (14%) dropped out before completing any of the web-based measures in the second part of the experiment, which is typical in web-based research [17], also contributed to the reduction of the data available for the comparison between lab and web-based testing in the present investigation. Therefore, our findings ~~need to~~ should be replicated in a larger study. A second limitation was that test-retest reliability was not examined here, given that the main aim of this study was to establish valid online versions of known individual difference measures. Future research should assess test-retest reliability, as it is as an interesting endeavor for studying individual difference measures in future work. Finally, and as indicated above, a third limitation concerned technical issues that affected data collection, as some participants used the wrong keys on the keyboard to submit their responses to the web-based version of the CVMT, rendering the data from some of the participants impossible to use for the comparison; furthermore, data from one subject was missing in the Web-based MLAT, which ~~could~~may have been due to technical ~~problems originated from~~issues at the participant's end (e.g., not following the general instructions given, such as refreshing or closing the browser page [see Procedure]; or Internet disconnection). In this sense, Reips and Krantz [61] (see also[17]) caution researchers that one of the potential disadvantages of Internet-driven testing is the technical variability characteristic of web-based research (e.g., different browsers and Internet connections), which, in turn, may affect data collection.

Conclusion

~~Despite the limitations, there are important contributions in this study.~~ 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 possible ~~possible~~ 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 and not comparing individual scores from one mode with scores from the other. Moreover, the extent to which the measures are equivalent varies according to the test. In this sense, we are confident that the two versions for the working memory test (OSpan) and the verbal declarative memory (MLAT5) are ~~fairly possibly measuring~~likely to measure the same construct, whereas the correlation ~~but we refrain from making such a strong statement~~ between~~for~~ the nonverbal declarative test (CVMT) versions was less pronounced~~, where the two modes might still plausibly measure strongly different aspects as well~~. Our research has shown that collecting experimentally controlled data on cognitive individual differences typically used in ~~L2~~ the area of L2 research in the Internet is feasible and comparable to lab-based collection. Consequently, some of these web-based versions could very well be incorporated, for example, in future web-based intervention studies on second language learning, thereby contributing to the scaling up of data collection in the field [~~50~~62-~~52~~64].

Acknowledgements

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

References

1. Kidd E, Donnelly S, Christiansen MH. Individual differences in language acquisition and processing. Trends in ~~cognitive~~ Cognitive ~~sciences~~Sciences. 2018;22(2):154-69. doi: 10.1016/j.tics.2017.11.006

2. 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. 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. 396-417.

~~4.~~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.

~~5.~~6. Larsen- Freeman D. Looking ahead: Future directions in, and future research into, second language acquisition. Foreign language annals. 2018;51(1):55-72. doi: 10.1111/flan.12314

~~6.~~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.

~~7.~~8. Lado B. Aptitude and pedagogical conditions in the early development of a nonprimary language. Applied Psycholinguistics. 2017;38(3):679-701.

~~8.~~9. 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.

~~9.~~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.

~~10.~~11. Buffington J, Morgan-Short K. Declarative and procedural memory as individual differences in second language aptitude. In: Wen Z, Skehan P, Biedroń A, Li S, Sparks R, editors. Language aptitude: Multiple perspectives and emerging trends. New York: Routledge; 2019. pp. 215–237.

~~11.~~12. 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.

~~12.~~13. Plonsky L. Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. Studies in Second Language Acquisition. 2013;35(4): 655-87. doi: 10.1017/S0272263113000399

~~13.~~14. Plonsky L. Quantitative research methods. In: Loewen S, Sato M, editors. The Routledge handbook of instructed second language acquisition. New York: Routledge; 2017. pp. 505-521.

15. Lindstromberg S. Inferential statistics in Language Teaching Research: A review and ways forward. Language Teaching Research. 2016;20(6): 741-68. doi: 10.1177/1362168816649979.

~~14.~~16. Tackett JL, Brandes CM, King KM, Markon KE. Psychology's replication crisis and clinical psychological science. Annual review of clinical psychology. 2019;15:579-604.

~~15.~~17. Krantz JH, Reips UD. The state of web-based research: A survey and call for inclusion in curricula. Behavior Research Methods. 2017;49(5): 1621-1619. doi: 10.3758/s13428-017-0882-x

18. Roever C. Web-based language testing. Language Learning & Technology. 2001;5(2):84-94.

19. Domínguez C, López-Cuadrado J, Armendariz A, Jaime A, Heras J, Pérez TA. Exploring the differences between low-stakes proctored and unproctored language testing using an Internet-based application. Computer Assisted Language Learning. 2019:1-27.

20. Diaz Maggioli GH. Web- Based Testing. The TESOL Encyclopedia of English Language Teaching. 2018:1-6.

21. Birnbaum MH. Human research and data collection via the Internet. Annu. Rev. Psychol.. 2004;55:803-32.

~~16.~~22. Hicks KL, Foster JL, Engle RW. Measuring working memory capacity on the web with the online working memory lab (the OWL). Journal of Applied Research in Memory and Cognition. 2016;5(4): 478-89. doi: 10.1016/j.jarmac.2016.07.010.

~~17.~~23. Wolfe CR. Twenty years of Internet-based research at SCiP: A discussion of surviving concepts and new methodologies. Behavior research methods. 2017;49(5): 1615-1620. doi: 10.3758/s13428-017-0858-x.

24. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. Value in Health. 2008;11(2): 322–333. doi: 10.1111/j.1524-4733.2007.00231.x.

25. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience. 2013;14(5):365.

26. Branch MN. The "Reproducibility Crisis:" Might the Methods Used Frequently in Behavior-Analysis Research Help?. Perspectives on Behavior Science. 2019;42(1):77-89.

27. Laraway S, Snycerski S, Pradhan S, Huitema BE. An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis. Perspectives on Behavior Science. 2019;42(1):33-57.

28. Shrout PE, Rodgers JL. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. Annual review of psychology. 2018;69:487-510.

29. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 2013.

~~18.~~30. Stewart N, Chandler J, Paolacci G. Crowdsourcing samples in cognitive science. Trends in cognitive sciences. 2017;21(10):736-48

~~19.~~31. Cowan N. Working memory maturation: Can we get at the essence of cognitive growth?. Perspectives on Psychological Science. 2016;11(2): 239-64. doi: 10.1177/1745691615621279

~~20.~~32. Baddeley AD. Modularity, working memory and language acquisition. Second Language Research. 2017;33(3): 299-311. doi: 10.1177/0267658317709852.

~~21.~~33. Roehr K. Linguistic and metalinguistic categories in second language learning. Cognitive Linguistics. 2008;19(1): 67-106. doi: 10.1515/COG.2008.005.

~~22.~~34. Grundy JG, Timmer K. Bilingualism and working memory capacity: A comprehensive meta-analysis. Second Language Research. 2017;33(3): 325-40. doi: 10.1177/0267658316678286.

~~23.~~35. Jeon EH, Yamashita J. L2 reading comprehension and its correlates: A meta-analysis. Language Learning. 2014;64(1):160-212. doi: 10.1111/lang.12034.

~~24.~~36. 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.

~~25.~~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/s13421-010-0034-3.

~~26.~~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.

~~27.~~39. 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.

~~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

~~29.~~41. 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.

~~30.~~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

~~31.~~43. Eichenbaum H. Hippocampus: cognitive processes and neural representations that underlie declarative memory. Neuron. 2004;44(1):109-20.

~~32.~~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.

~~33.~~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.

~~34.~~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.

~~35.~~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.

~~36.~~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.

~~37.~~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.

~~38.~~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.

~~39.~~51. Carroll JB, Sapon SM. Modern Language Aptitude Test: Manual. New York: Psychological Corporation; 1959.

~~40.~~52. Trahan DE, Larrabee GJ. Continuous visual memory test. Odessa, FL: Assessment Resources. 1988.

~~41.~~53. Schneider W, Eschman A, Zuccolotto A. EPrime user’s guide. Pittsburgh, PA: 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.

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.

~~43.~~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. doi: 10.1037/0096-3445.133.2.189.

~~44.~~

~~45. 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.~~

~~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. 1;68(2): 321-391. doi: 10.1111/lang.12286.~~

~~47.~~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.

~~48.~~60. Leidheiser W, Branyon J, Baldwin N, Pak R, McLaughlin A. Lessons learned in adapting a lab-based measure of working memory capacity for the web. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2015. Los Angeles: Sage CA; 2015. pp. 756-760.

61. Reips UD, Krantz JH. Conducting true experiments on the Web. In: Gosling SD, Johnson JA, editors. Advanced methods for conducting online behavioral research. Washington, DC: American Psychological Association; 2010. pp. 193-216.

~~49.~~62. MacWhinney B. A shared platform for studying second language acquisition. Language Learning. 2017;67(S1): 254-75. doi: 10.1111/lang.12220.

~~50.~~63. Meurers D, Dickinson M. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. Language Learning. 2017;67(S1): 66-95. doi: 10.1111/lang.12233.

~~51.~~64. Ziegler N, Meurers D, Rebuschat P, Ruiz S, Moreno- Vega JL, Chinkina M, Li W, Grey S. Interdisciplinary research at the intersection of CALL, NLP, and SLA: Methodological implications from an input enhancement project. Language Learning. 2017;67(S1): 209-231. doi: 10.1111/lang.12227.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

LEAD
Graduate School &
Research Network

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, 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, l.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.*

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

LEAD Graduate School & Research Network

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