Individual differences in online research: Comparing lab-based and online administration of a psycholinguistic battery of linguistic and domain-general skills

Kyla McConnell (1, 2), Florian Hintz (1,3) and Antje S. Meyer (1,2)

1 Max Planck Institute for Psycholinguistics, Nijmegen

2 Donders Centre for Brain, Cognition, and Behavior, Radboud University, Nijmegen

3 Philipps University, Marburg (DE)

Corresponding author:
Kyla McConnell
kyla.mcconnell@mpi.nl
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD, Nijmegen
The Netherlands
+31 243521379

Abstract

Experimental psychologists and psycholinguists increasingly turn to online research for data collection, due to the ease of sampling a large number of diverse participants in parallel. Online research has shown promising validity and consistency, but is it suitable for all paradigms? Specifically, is it reliable enough for individual difference research? The current paper reports performance on fifteen tasks from a psycholinguistic individual difference battery, including timed and untimed assessments of linguistic abilities, as well as domain-general skills. From a demographically homogenous sample of Dutch young people, 149 participants took part in the study in the lab and 515 participated online. Our results indicate that there is no reason to assume that participants tested online will underperform compared to lab-based testing, though they highlight the importance of motivation as well as the potential for external help (e.g. through looking up answers) online. Overall, we conclude that there is reason for optimism in the future of online research into individual differences. (159 words)

Keywords: individual differences, online experimentation, psycholinguistics, domain-general skills, language skills

1. Introduction

Over the last several years, experimental psychologists and psycholinguists have begun to explore the potential of conducting studies online, outside of controlled laboratory environments. In particular, the development towards increased online research was accelerated by the restrictions on lab-based work during the height of the Covid-19 pandemic, which forced many researchers to explore whether an online lab could be brought to the participants' home. Researchers quickly began to share resources about best practices as well as technical and practical recommendations for behavioral experimentation outside of the lab (Blake & Dąbrowska, 2024; Garcia et al., 2022; Grootswagers, 2020; Rodd, 2024; Sauter et al., 2020).

But do results collected online really compare with those collected in laboratory environments? Large, stable, and group-level effects seem to replicate well online. However, there is an increasing interest in psychology and psycholinguistics in differences between individuals and their causes (Dąbrowska, 2012; Engelhardt et al., 2017; Isbilen et al., 2022; Kidd et al., 2018, 2023; McConnell, 2023; McConnell & Blumenthal-Dramé, 2021; Payne et al., 2014; Pronk et al., 2022). Yet individual difference paradigms may face unique challenges when moved to an online lab. First of all, they require precise estimation at the participant level (Hedge et al., 2018). Noise that may be leveled out in factorial designs could lead to erroneous conclusions about individual performance. And of course, systematic differences in participant motivation, or any other extraneous pressures on participants, could drive or obscure effects. It is not necessarily the case that paradigms that produce sensible data at the group level under online data collection will also be reliable in an individual difference paradigm (Haines et al., 2020; Hedge et al., 2018).

To address the suitability of online data collection to individual difference research in psycholinguistics, we compared performance on an extensive individual difference battery completed by large, demographically similar samples in a controlled lab setting or online. Results suggest reason for optimism in the use of online data collection for measuring stable individual differences; however, the role of two important considerations (looking up answers and motivation level) arise as important factors to keep in mind.

1.1 Online research: Potentials and pitfalls

Psycholinguistic labs exist for good reasons. They ensure that all participants are tested under identical conditions, which are optimized for the purpose of the experiment (e.g. in a quiet, well-lit room without much distraction), and that they all fully understand what they are meant to do. Some

details hardly need to be considered in lab research, such as the time of day (which influences visual perception through lighting, as well as participant tiredness,) but can vary widely when research is taken out of the lab (Dandurand et al., 2008).

Despite the less controlled nature of online testing, copious previous research has established its validity for experimental psychology. Many classic psychological effects replicate well under online data collection, including psychometric assessments. For instance, comparable performance has been found in measures of cognitive inhibition like the Stroop, Flanker, Simon and go/no-go tasks, measures of memory like digit spans and two-back tasks, as well as visual search and attentional blink paradigms (Crump et al., 2013; Germine et al., 2012; Miller et al., 2018; Semmelmann & Weigelt, 2017). Challenging higher-order cognitive tasks such as face reading and memory for abstract art also pattern similarly in online compared to lab-based data (Germine et al., 2012). Standard psycholinguistic effects have also been replicated in online studies, including effects of word frequency, age of acquisition and name agreement on the speed of picture naming, lexical decision, and self-paced reading, among others (Corps & Meyer, 2023; Fairs & Strijkers, 2021; He et al., 2021; Hilbig, 2016).

Importantly, results are also similar when the same participants complete the same tasks both online and in the lab. Miller and colleagues (2018) invited 127 participants to complete multiple tasks assessing processing speed, including the go/no-go task, the two-back task, and a number-letter task, both in the lab and online, with one week between the two settings. A diffusion model showed no significant difference in central tendency, and the internal reliability of the test scores was similar for the two sessions. Slightly more variance (approximately 5%) was observed in online administration, however, and test-retest reliability between the two sessions was weak for some tasks (range: .33 to .73). Overall, the results strongly suggest that online data collection is comparable to data in the lab. However, it is hard to say if this conclusion extends to individual difference designs because the data were analyzed at the group level.

Of course, collecting research data online is not without risk. Two primary issues can be identified: technical concerns related to the quality of the experimental software and the hardware available to the participants, and issues related to participants' behavior in a remote setting. The first of these issues has been addressed for many of the most common experimental platforms and programming environments (Anwyl-Irvine et al., 2021; Bridges et al., 2020; Reimers & Stewart, 2015), and is not considered here. Instead we focus on the second issue, i.e. how participants' behavior is affected by the testing environment. Specifically, we compare the scores of young Dutch students tested online or in the lab. A comparison of demographically more varied groups (e.g. students vs. online

crowdsourced workers) is beyond the scope of the current paper (but see Hauser & Schwarz, 2016; Peer et al., 2021). In this, our study provides novel information about the way test scores may differ between relatively homogeneous well-matched groups when tests are administered in a lab or online.

A major factor in participant behavior is likely to be motivation. In psycholinguistics, there is growing appreciation for the role of motivation (or the lack thereof) in language tasks (Christianson et al., 2022). This includes the realization that "... some portion of published work [consists] of data that have been collected from unmotivated, uninterested, or disengaged participants", even if it is collected in the lab (Christianson et al., 2022, p. 54). We might expect that participants tested online are overall less motivated or struggle to maintain an appropriate level of motivation for the duration of a study. In comparison, participants in the lab have already committed to coming to a new environment and are participating under the careful eye of an experimenter. Indeed, online data is often found to contain more variation, though the exact causes for this have not been identified (Miller et al., 2018; Semmelmann & Weigelt, 2017).

Motivation may have an effect on the task level, in that certain task types might demand more attention and focus than others. Difficult tasks see considerable rates of drop-out in online experiments, though an engaging online test-taking environment can mitigate the difference in performance (Dandurand et al., 2008; Pedersen et al., 2023; Rodd, 2024). This may be particularly relevant to individual difference studies if drop-out is self-selecting; that is, if participants are more likely to withdraw from an online experiment if they find it particularly difficult (e.g. because they have a weakness in the tested skill or because they are less able to resist distraction). Drop-out is likely to be less severe in in-person settings because participants would have to excuse themselves from the room and leave the building.

Additionally, participants may not be willing to create conditions that would facilitate performance in online experimentation. For instance, only half of participants in one online study reported doing the experiment in a quiet environment (Simcox & Fiez, 2014). In another study, nearly a third of online participants reported watching TV while doing experimental tasks (Chandler et al., 2014). Applied to an individual difference context, participants who are more prone to distraction, less motivated, or otherwise less engaged may be especially likely to multitask, thus lowering their scores. Again, in the laboratory, these variables are the researcher's responsibility, and not within participant's control.

On the other hand, certain participants may strive too strongly for peak performance. Online participants have the opportunity to score higher than their in-lab counterparts in any task where

they can easily look up answers. For some tasks, this is as easy as opening another browser window, glancing at one's phone, or even asking a nearby friend for a second opinion. In one study, 10% of crowdsourced online participants correctly indicated the number of countries in Africa, compared to none in a face-to-face sample (Goodman et al., 2013). Participants can possibly be discouraged from using external help by clear instructions and reassurance that they are not expected to get every question correct, but it is difficult to entirely eliminate such behavior (Goodman et al., 2013). Perhaps this is simply linked to modern life, where any information is merely a click away.

Importantly, a bulk of the previous research on the difference between research in the lab and on the web focuses on group-level differences. Yet online research is especially advantageous to individual differences work. Correlational research demands large sample sizes with sufficient within-sample variability, and these two aspects can be difficult to obtain through convenience sampling in the (university) population close to the lab. In contrast, online experiments have the potential to reach participants that are representative of the general public and many participants can be tested in parallel (Berinsky et al., 2012). Yet the questions raised about data quality, the ability to seek external help, and the role of motivation mean that a comparison of similar participants' performance in the lab and online is still needed. To assess these questions, we compare data from fifteen different linguistic and non-linguistic tasks collected either online or in the lab as part of the Individual Differences in Language Skills (IDLaS-NL) battery.

1.3 IDLaS-NL, online and in the lab

The data from the IDLaS-NL battery is ideal for investigating the feasibility of individual differences research in an online environment. The battery was designed for young adult speakers of Dutch and includes 31 different assessments of linguistic and non-linguistic skills (Hintz et al., 2023)[1]. The full test battery was completed online over four sessions by a sample of 579 young Dutch speakers. In addition, 169 participants completed the battery in a designated lab space under the supervision of an experimenter.

Online participants completed all 31 tasks of the battery using the newly-developed experimental software *Frinex* (FRramework for INteractive EXperiments). For inhouse testing, 15 tasks were implemented using *Frinex*, while the other 16 used the software *Presentation*, which is designed for laboratory and not online use. We compared only the fifteen tasks that both groups completed on

[1] IDLaS-NL is also a customizable battery that can be used by any interested researcher to create and run an individual difference battery for speakers of Dutch. For more information, see: www.mpi.nl/idlas-nl and Hintz et al. (2023).

Frinex: five untimed assessments of receptive linguistic skills or linguistic experience, three timed speech production tasks, four assessments of working memory and an assessment of nonverbal reasoning. These fifteen tasks thus cover both language comprehension and production, as well as domain-general cognitive ability. In the following section, each of these tests will be described briefly and functionally; for further details about the technical setup and the design of each task, the reader is directed to Hintz et al. (2020)[2].

This dataset has several advantages when considering the current question. First of all, the sample size is large, which affords strong statistical power. At the same time, the participants all come from the same, rather homogenous population: Dutch-speaking young people (primarily students) between the ages of 18 and 30. Further, we administered a large battery of tests from different domains of linguistic and non-linguistic ability. This allows us to assess the impact of online testing in general, but also the relative impact of the testing environment on different test types. For example, some types of assessments may be easier to look up the answers to, while others demand more motivation or focus on the part of the participant.

In summary, the IDLaS-NL dataset contains data from a large sample of demographically similar participants on several linguistic and domain-general aspects of individual difference. All participants completed the tasks in the same technological set-up. In the following section, we outline the tasks involved in the battery, and look at their distributions under online or laboratory assessment. We also submit them to a hierarchical Bayesian model to ensure that the trends seen in the descriptive statistics are maintained. In doing so, we assess the effect of online testing on fifteen diverse tasks from the domains of language production, language comprehension and domain-general processes. We ask how participant-based factors such as motivation and the potential to look up answers affect data collected online. Ultimately, we aim to determine whether online testing is suitable for individual differences designs, across a range of tasks typical to behavioral psycholinguistics.

2. Method

2.1 Participants

169 participants completed the test battery in a designated psycholinguistic lab at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands. Participation was spread over two days, with two hour-long blocks on each day. Participants were tested individually or in groups of up to

[2] Test validity statistics are additionally available here: https://www.mpi.nl/idlas-nl/faq

four persons. 579 participants took part in the same tasks online, using a simulated version of a Chrome browser. Only the fifteen tasks that were performed on the same experimental software (Frinex) are considered here. A full description of the lab setup can be found in Hintz et al. (in prep).

In the current paper, we only consider participants that had usable scores for all fifteen tasks. In the lab sample, 20 participants did not have full datasets (12%) and in the online sample, 64 did not have full data sets (11%). This left data from 149 participants in the lab and 515 online. Table 1 shows which tasks contained missing values. Note that some participants were missing more than one task. Overall, we see that despite the fact that the percentage of participants without a full dataset is very similar in the two samples, there is more data missing from the language production tasks online than in the lab. In particular, for the maximum speech rate task (reciting the months of the year as fast as possible), this was often because participants mumbled or dropped syllables, leading to trials that could not be used. As for the verbal fluency tasks, scores were discarded when participants misunderstood the directions (e.g. by listing only incorrect words). This did not happen often in the lab, where participants could ask questions about the instructions.

Task	in-lab	online
Antonym Production	0	0
Idiom Recognition	0	0
Spelling	0	1
Author Recognition (DART)	3	1
Prescriptive Grammar	0	0
Peabody (PPVT)	0	1
Digit span (forward)	0	2
Digit span (backward)	1	3
Corsi span (forward)	3	2
Corsi span (backward)	1	2
Nonverbal Reasoning	3	0
Verbal Fluency (semantic)	3	16
Verbal Fluency (phonological)	1	17
RAN	3	6
Max Speech Rate	5	24

Table 1: Number of participants missing a score, by task and setting

Table 2 shows the number and percentage of participants based on gender, whether they were a student and what their highest achieved level of education was at the time of testing. The Dutch education system has been simplified here, so that the label “High school” includes the qualifications *HAVO*, *VWO*, *(V)MBO*, *Basisonderwijs*, and anyone who selected “other” (N = 5). Participant ages are visualized in Figure 1. Although the values differ slightly, the samples are similar; participants were all Dutch young people between the age of 18 and 30, largely college-educated or in a university program. The most noticeable difference is that the online testing had more participants who identified as female (78.6% online compared to 53.7% in the lab). Although the age distributions vary slightly, the mean age in both settings is comparable (22.4 in the lab, 22.9 online).

N	Female	Male	Student	High School	Bachelor	Master
in-lab						
149	80 (53.7%)	68 (45.6%)	119 (79.9%)	88 (59.1%)	48 (32.2%)	13 (8.7%)
online						
515	405 (78.6%)	108 (21.0%)	379 (73.6%)	226 (43.9%)	204 (39.6%)	85 (16.5%)

Table 2: Demographic characteristics of in-lab and online participants, respectively.

Figure 1: Participant ages for participants in the lab and online, respectively. The boxplots show the median, interquartile range (IQR) and range and are overlaid on (mirrored) histograms to show the distribution of the data.

2.2 Untimed linguistic tasks

The untimed linguistic tasks included in this analysis are an antonym production test, the Dutch Author Recognition Test (DART), an idiom recognition task, the Peabody Picture Vocabulary Test, a prescriptive grammar task and a spelling task. All of these assessments tap into the participant's linguistic knowledge or experience in an untimed test setting. Each test is described briefly in terms of the participant's task below. The descriptive statistics for each sample for each test are shown in Table 3 and their distributions are visualized in Figure 2.

Antonym production: For this task, participants both saw and heard a word, then were asked to say the antonym to this word (Mainz et al., 2017). There were 25 trials. The task was not timed, and the participant's score was the proportion of correctly produced antonyms.

Dutch Author Recognition Test (DART): Participants completed the Dutch Author Recognition Test (DART) (Brysbaert et al., 2020), which is a common assessment of exposure to print language. For this, they saw 132 names, of which 90 were real authors and 42 were non-author foils. Their task was to identify which of the names were authors. A participant's score was the proportion of correctly identified authors minus the proportion of 'false alarms' (foils that the participant indicated were authors).

Idiom recognition: Participants saw 10 Dutch idioms selected from a normed database (Hubers et al., 2019). One idiom appeared on the screen at a time, together with four possible meanings; both idioms and meanings could be listened to by clicking on an icon. The task was to select the correct meaning for the idiom, and the score was calculated as the proportion of trials for which the correct meaning was selected.

Peabody Picture Vocabulary Test (PPVT): Participants completed the online version of the standardized Dutch-language PPVT (third edition) (Dunn & Dunn, 1997; Schlichting, 2005). On each trial, they saw four drawings and heard a single word. Their task was to select which of the four pictures best matched the meaning of the word they heard. The assessment has a standard staircase procedure, so that participants moved to a harder block (of 12 items) only if they performed well enough in the previous block; the test was started at the block recommended for people aged 18 to 35. A participant's score was calculated, as recommended, as the number of the last item they responded to, minus the amount of errors they had made in the test.

Prescriptive grammar: For this task, participants heard 40 sentences and had to indicate whether they were grammatically correct. Half of the stimuli set featured correct Dutch sentences, while the other half had one of five common errors: the incorrect use of a personal pronoun ("ze", they vs. "hun", their; "ik", I vs. "mij", me), a comparative ("als", as vs. "dan", than), a relative pronoun ("die", this vs. "dat", that) and the participles of complex verbs (e.g., "stofgezogen", vacuumed). A participant's score was the proportion of correct responses.

Spelling: The spelling assessment contained 60 words that are frequently misspelled in Dutch. Half the items were spelled correctly and half represented common misspellings. Participants saw all words at once and were asked to identify which words were spelled incorrectly. A participant's score was calculated as the proportion of correctly identified misspelled words minus the proportion of 'false alarms' (words spelled correctly that the participant indicated were misspelled).

Untimed linguistic tasks					
Setting	**Mean**	**Median**	**SD**	**Min**	**Max**
Antonym Production					
In-lab	0.80	0.80	0.07	0.60	0.96
Online	0.79	0.80	0.08	0.56	0.96
Author Recognition (DART)					
In-lab	0.27	0.23	0.14	0.02	0.80
Online	0.29	0.26	0.15	0.00	0.81
Idiom Recognition					
In-lab	0.78	0.80	0.11	0.40	1.00
Online	0.77	0.80	0.13	0.40	1.00
Peabody (PPVT)					
In-lab	176.10	177.00	9.32	148.00	198.00
Online	176.79	178.00	10.36	141.00	198.00
Prescriptive Grammar					
In-lab	0.69	0.70	0.10	0.42	0.92
Online	0.70	0.70	0.11	0.38	0.98
Spelling					
In-lab	0.59	0.60	0.16	0.10	0.93
Online	0.63	0.63	0.15	0.13	1.00

Table 3: Descriptive statistics (mean, median, SD, min and max) for all untimed linguistic tasks, across in-lab and online samples

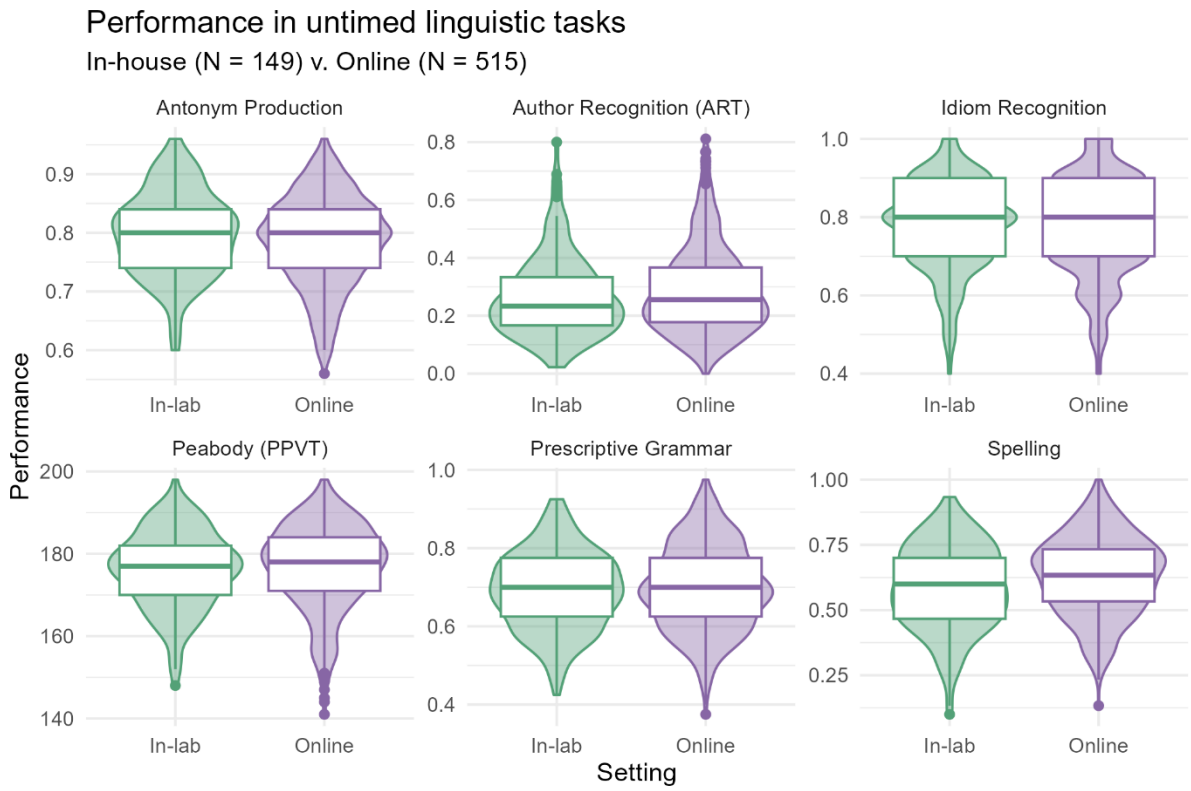

Figure 2: Distribution of performance in the untimed linguistic tasks, across in-lab and online samples. The boxplots show the median, interquartile range (IQR) and range and are overlaid on (mirrored) histograms to show the distribution of the data.

2.3 Timed speech production tasks

The second category of tasks that we analyzed are the timed speech production tasks. For this domain, we had two verbal fluency tasks, one of which was phonological and the other was semantic. We also included the maximum speech rate and rapid automatized naming (RAN). The tasks are described briefly, their descriptive statistics are listed in Table 4 and their distributions are shown in Figure 3.

Maximal speech rate: Participants were asked to name the months of the year as quickly as possible, while maintaining good pronunciation. There were two identical trials. The score was calculated as the average duration (measured from utterance onset to offset) over both trials, if both were correct; otherwise, the duration of the correct trial was used. The duration was log-transformed and inversed.

Rapid Automatized Naming (RAN): This test is a classic assessment of speeded word form access (Denckla & Rudel, 1976). Participants were familiarized with five line drawings and a name for each

of these items, which was the most common name in a norming study (Araújo et al., 2021). They then saw all the drawings arranged in 5x6 grid and their task was to name all objects from left to right, starting at the top row and proceeding to the following rows. There were four sets featuring different pictures. Each set was used twice in different orders within the grid. The participant's score was the ratio of the correctly named objects by the duration of speech (in seconds) during the trial, averaged over all sets.

Verbal fluency (semantic): Participants were given a semantic category (for the first trial, "animals", and for the second, "food and drink"). They were given one minute to name as many words that fit the category as possible (Shao et al., 2014). Their score was the number of unique words produced in a minute, averaged over both categories.

Verbal fluency (phonological): Participants were given a letter (for the first trial, "M", and for the second, "S"). They were given one minute to name as many words that started with that letter as possible (Shao et al., 2014).Their score was the number of unique words produced in a minute, averaged over both letters.

Timed linguistic production tasks					
Setting	**Mean**	**Median**	**SD**	**Min**	**Max**
Max Speech Rate					
In-lab	-3.67	-3.66	0.09	-3.94	-3.45
Online	-3.66	-3.65	0.09	-4.10	-3.44
Rapid Automatized Naming (RAN)					
In-lab	1.60	1.58	0.23	1.00	2.29
Online	1.59	1.59	0.24	0.88	2.27
Verbal Fluency (semantic)					
In-lab	24.87	24.50	4.48	16.50	38.50
Online	25.36	25.50	4.89	10.00	39.50
Verbal Fluency (phonological)					
In-lab	15.23	15.50	4.12	6.00	24.00
Online	15.76	15.50	4.35	5.00	32.00

Table 4: Descriptive statistics (mean, median, SD, min and max) for all timed linguistic production tasks, across in-lab and online samples.

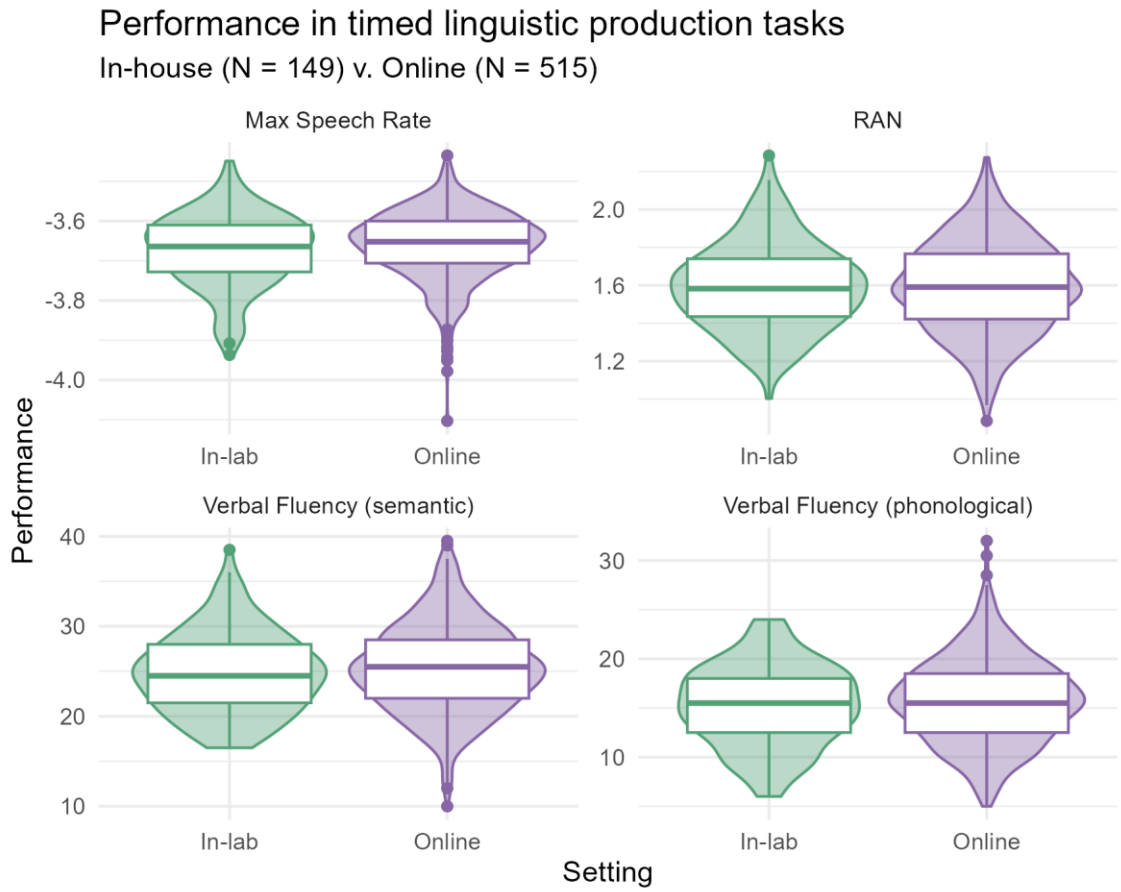

Figure 3: Distribution of performance in the timed linguistic production tasks, across in-lab and online samples. The boxplots show the median, interquartile range (IQR) and range and are overlaid on (mirrored) histograms to show the distribution of the data.

2.4 Domain-general skills

There were five assessments of domain-general skills, four of which tested working memory. The Corsi block clicking task measures visual working memory while the digit span measures verbal working memory. We also assessed nonverbal reasoning via Raven's Advanced Progressive Matrices. The descriptive statistics for these tasks are listed in Table 5 and visualized in Figure 4.

Corsi span (forward and backward): The Corsi block clicking test (Berch et al., 1998) assesses visuospatial short-term memory. Participants saw nine blocks on the screen, which lit up one by one at a speed of one square per second. Their task was to recall this sequence and re-create it after the prompt ended by clicking on the blocks. In the forward version, they clicked on the blocks in the same order as they saw them light up in the prompt phase. In the backward version, they clicked on them in the reverse order. Trials started at a length of three consecutive squares (after the practice

items), and participants saw a longer sequence if they answered one of two trials of the same sequence length correctly. That is, if they correctly answered one of two trial in which three squares were lit, the next trials would have four squares light up. The assessment ended when two consecutive trials were answered incorrectly, or at the end of the test (spans of nine squares). A participant's score was the sum of correct responses over the version (forward or backward).

Digit span (forward and backward): Participants heard spoken numbers, and were asked to type the numbers they heard into a text box after auditory playback had completed (Wechsler, 2004). For the forward version, they typed numbers in the same order they had heard them, and for the backward version, in the opposite order (starting with the last number they heard). Similar to the Corsi span, trials started at 3 digits long (after the practice items), and participants heard a sequence that was longer by one digit if they answered two consecutive trials correctly. The assessment ended when two consecutive trials were answered incorrectly, or at the end of the test (spans of 9 digits in the forward version, and 8 in the backward version). A participant's score was the sum of correct responses over the version.

Nonverbal reasoning: Nonverbal reasoning was assessed using the Raven's Advanced Progressive Matrices test (Raven et al., 1998). For this, participants saw a cue and were asked to pick one of eight shapes/visual items that best completed the pattern shown in the cue. Participants could skip items and return to them later, or indicate that they did not know the answer. There were 36 items (excluding practice trials) and participants had 20 minutes to answer as many as possible. Their score was calculated as the proportion of trials they answered correctly.

Domain-general skills					
Setting	**Mean**	**Median**	**SD**	**Min**	**Max**
Corsi span (backward)					
In-lab	7.73	8.00	2.05	2.00	14.00
Online	7.65	8.00	2.05	1.00	14.00
Corsi span (forward)					
In-lab	8.48	8.00	1.77	4.00	13.00
Online	8.29	8.00	1.90	1.00	14.00
Digit span (backward)					
In-lab	7.28	7.00	2.09	2.00	12.00
Online	7.38	7.00	2.22	2.00	12.00
Digit span (forward)					
In-lab	8.48	8.00	2.10	4.00	14.00
Online	8.88	9.00	2.11	4.00	14.00
Non-verbal reasoning (Raven's)					
In-lab	0.60	0.61	0.15	0.22	0.94
Online	0.57	0.58	0.14	0.06	0.92

Table 5: Descriptive statistics (mean, median, SD, min and max) for all domain-general tasks, across in-lab and online samples.

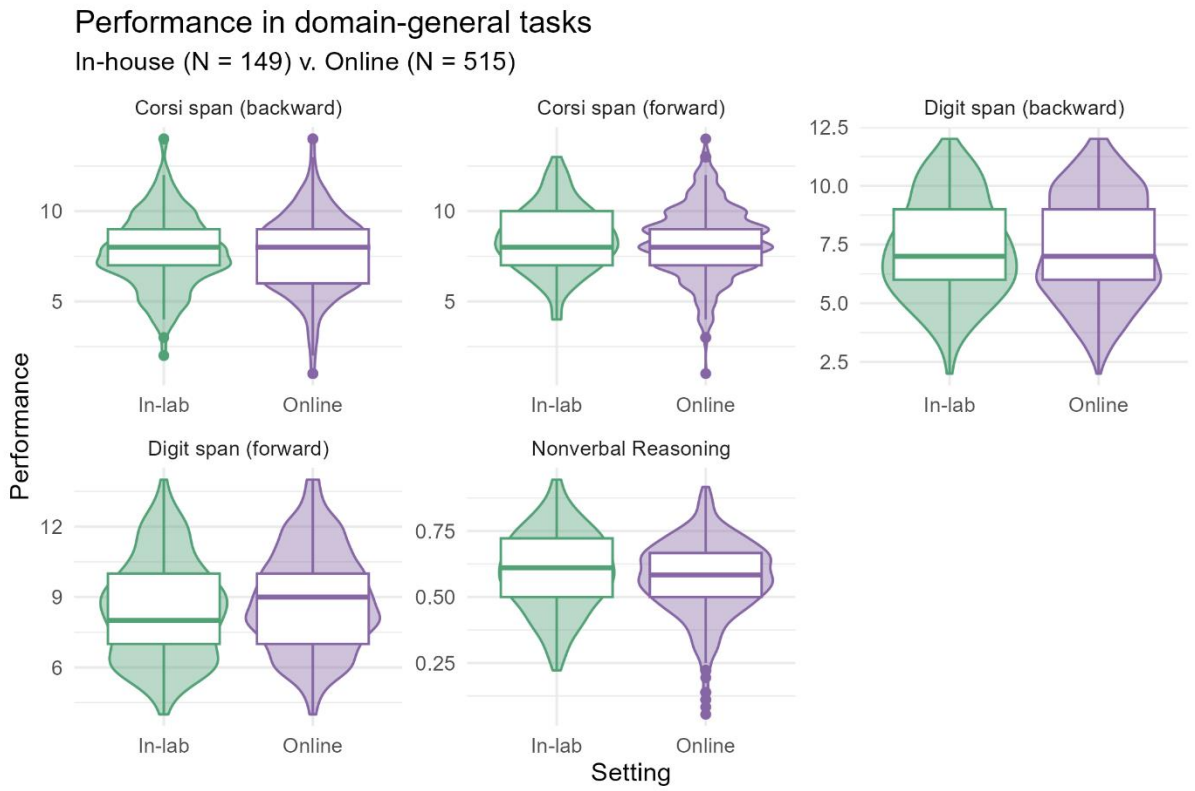

Figure 4: Distribution of performance in the domain-general tasks, across in-lab and online samples. The boxplots show the median, interquartile range (IQR) and range and are overlaid on (mirrored) histograms to show the distribution of the data.

3. Results

Looking at the raw data and descriptive statistics, we can see that performance online was very similar to performance in the lab for the majority of tasks. Both the summary and the distributions visualized with violin and box plots are also highly comparable. Nonetheless, performance across the two groups was also assessed in a Bayesian mixed model. This was a useful inferential tool because the question at hand revolves around support for the null hypothesis (rather than, for example, submitting the data to a frequentist model, which cannot support the null, just quantify evidence against it.) Further, the by-subject random intercepts control for any systematic by-participant variance that may differ between the two groups. We describe the model parameters and output in the following section.

3.1 Preprocessing

In the current analysis, we compare fifteen tasks on very different scales; some are measures of accuracy in proportion to the total trials, others represent the number of words recalled, or the duration of articulation. In order to compare performance on the diverse tasks, we assess a participant's performance on a task in relationship to the mean and the standard deviation of scores across all participants in a given task. For this, we calculated z-scores by task, so that a value of 0 represents the mean performance in the given task, and a value of 1 or -1 represents performance one standard deviation above or below the mean, respectively. This z-score served as the dependent variable for the models described below.

3.2 Model

To statistically evaluate the role of setting (in-lab or online) of each of the tasks, we fit a Bayesian mixed model predicting a participant's performance by the interaction between task and setting and a random intercept by participant, using the R package *brms* (Bürkner, 2017). A main effect of task was not predicted, since the dependent variable was z-scored by task (i.e., the mean of every task was 0, and the SD was 1). A main effect of setting was also uninteresting because it would have abstracted across all fifteen tasks, not all of which we assumed to have the same relationship to setting (i.e. some tasks may be faster online where others are slower.) Thus, the model formula was always: Performance ~ Task:Setting + (1|Participant). Both Setting (2 levels) and Task (14 levels) were sum coded to facilitate the interpretation of the interaction term.

Priors were centered at 0, since we could expect setting and task to affect the scores either positively or negatively. Because the DV is a z-score with mean 0 and SD 1, the prior for the intercept was set relatively straight-forwardly to be quite certain around 0 (a normal distribution with mean 0 and SD 0.5), and we set the same prior for the residual error sigma, which is automatically trimmed by brms to not allow values below zero. Because the design was optimized to capture differences between individuals, we set a slightly wide prior for the by-participant random intercept, a normal distribution with mean 0 and SD 1.

For the interaction of task and setting, we set three different priors, so that we could later check that the prior wasn't exerting undue pressure on the results. For this, we fit:
(a) a "wide" prior – a normal distribution with mean 0 and SD 1;
(b) a "moderate" prior – a normal distribution with mean 0 and 0.5;
(c) and a "narrow" prior – a normal distribution with mean 0 and SD 0.2.

Priors were investigated visually before model fitting to ensure that they were reasonable in scale. We focus primarily on the model with "moderate" priors; the results for the other two models are very similar, and are available together with the full analysis script and data: https://osf.io/2knzc/

Task	Setting	Estimate	Std. Error	CI (low)	CI (high)
Intercept	NA	0.0032	0.0946	-0.1809	0.1892
By-participant random intercept	NA	0.4260	0.0155	0.3964	0.4569
Residual by-item variance	NA	0.9193	0.0068	0.9061	0.9327
Antonym Production	In-lab	0.0770	0.1185	-0.1562	0.3066
Antonym Production	Online	-0.0845	0.1014	-0.2817	0.1126
Author Recognition	In-lab	-0.0789	0.1187	-0.3115	0.1538
Author Recognition	Online	0.0742	0.1019	-0.1252	0.2714
Digit Span (backward)	In-lab	-0.0244	0.1189	-0.2588	0.2066
Digit Span (backward)	Online	0.0196	0.1019	-0.1812	0.2173
Digit Span (forward)	In-lab	-0.0948	0.1180	-0.3257	0.1356
Digit Span (forward)	Online	0.0905	0.1022	-0.1102	0.2894
Idiom Recognition	In-lab	0.0232	0.1187	-0.2096	0.2560
Idiom Recognition	Online	-0.0296	0.1020	-0.2292	0.1686
Max Speech Rate	In-lab	-0.0642	0.1184	-0.2949	0.1670
Max Speech Rate	Online	0.0594	0.1019	-0.1405	0.2571
Nonverbal Reasoning	In-lab	0.0848	0.1191	-0.1507	0.3183
Nonverbal Reasoning	Online	-0.0915	0.1021	-0.2925	0.1050
Picture Vocabulary	In-lab	-0.0366	0.1189	-0.2718	0.1938
Picture Vocabulary	Online	0.0308	0.1022	-0.1713	0.2282
Prescriptive Grammar	In-lab	-0.0385	0.1184	-0.2706	0.1932
Prescriptive Grammar	Online	0.0335	0.1017	-0.1664	0.2303
Rapid Automatized Naming	In-lab	0.0020	0.1184	-0.2299	0.2361
Rapid Automatized Naming	Online	-0.0076	0.1021	-0.2092	0.1888
Spelling	In-lab	-0.1426	0.1184	-0.3778	0.0871
Spelling	Online	0.1390	0.1015	-0.0605	0.3347
Verbal Fluency (phon)	In-lab	-0.0636	0.1192	-0.2950	0.1701
Verbal Fluency (phon)	Online	0.0576	0.1021	-0.1437	0.2551
Verbal Fluency (sem)	In-lab	-0.0537	0.1183	-0.2874	0.1751
Verbal Fluency (sem)	Online	0.0490	0.1023	-0.1513	0.2467
Visual Span (forward)	In-lab	0.0174	0.1185	-0.2141	0.2475
Visual Span (forward)	Online	-0.0237	0.1023	-0.2236	0.1744
Visual Span (backward)	In-lab	0.0472	0.1184	-0.1869	0.2798
Visual Span (backward)	Online	-0.0538	0.1022	-0.2544	0.1430

Table 6: Model output from Bayesian mixed model predicting z-score performance by the interaction between Task and Setting, from the model with a "moderate" prior for beta.

Looking at the numeric model output in Table 6, we first see that the intercept is, as expected, estimated at 0.00. Most of the tasks differ by Setting to only a very minimal degree, and all credible intervals cross 0. The by-setting estimates should be interpreted as the difference between the grand mean and the setting, not the raw difference between the two settings. All estimates of the interaction between task and setting have credible intervals contain 0, showing that the true effect cannot be distinguished from 0. This indicates that there is no evidence for differences across the two types of test administration.

Despite this, there are a few tasks worth noticing. In the Spelling task, participants online have better scores (b = 0.139 [-0.0605, 0.3347]) compared to in the lab (b = -0.1426 [-0.3778, 0.0871]). Scores for the forward digit span are also better online (b = 0.0905 [-0.1102, 0.2894]), though the scores from the backward digit span does not show the same pattern (b = 0.0196 [-0.1812, 0.2173]). On the other hand, for Nonverbal Reasoning, participants in-lab have better scores (b = 0.0848 [-0.1507, 0.3183]) compared to online (b = -0.0915 [-0.2925, 0.105]). Antonym Production also shows a similar pattern, in that participants in the lab perform better (b = 0.077 [-0.1562, 0.3066]) compared to online (b = -0.0845 [-0.2817, 0.1126]). However, credible intervals for all of these terms cross 0.

We also see that participants vary significantly compared to each other. The estimate for by-participant variation is (b = 0.426 [0.3964, 0.4569]). There is also considerable by-item variation, resulting in residual error (b = 0.9193 [0.9061, 0.9327]); this is difficult to interpret, however, as the items come from different combinations of participants and items. The conditional effects are visualized in Figure 7. The figure makes clear that the estimates for the in-lab setting are less exact, because this group is smaller. We also see that the by-setting estimates for the spelling task do not overlap, unlike those for all other tasks.

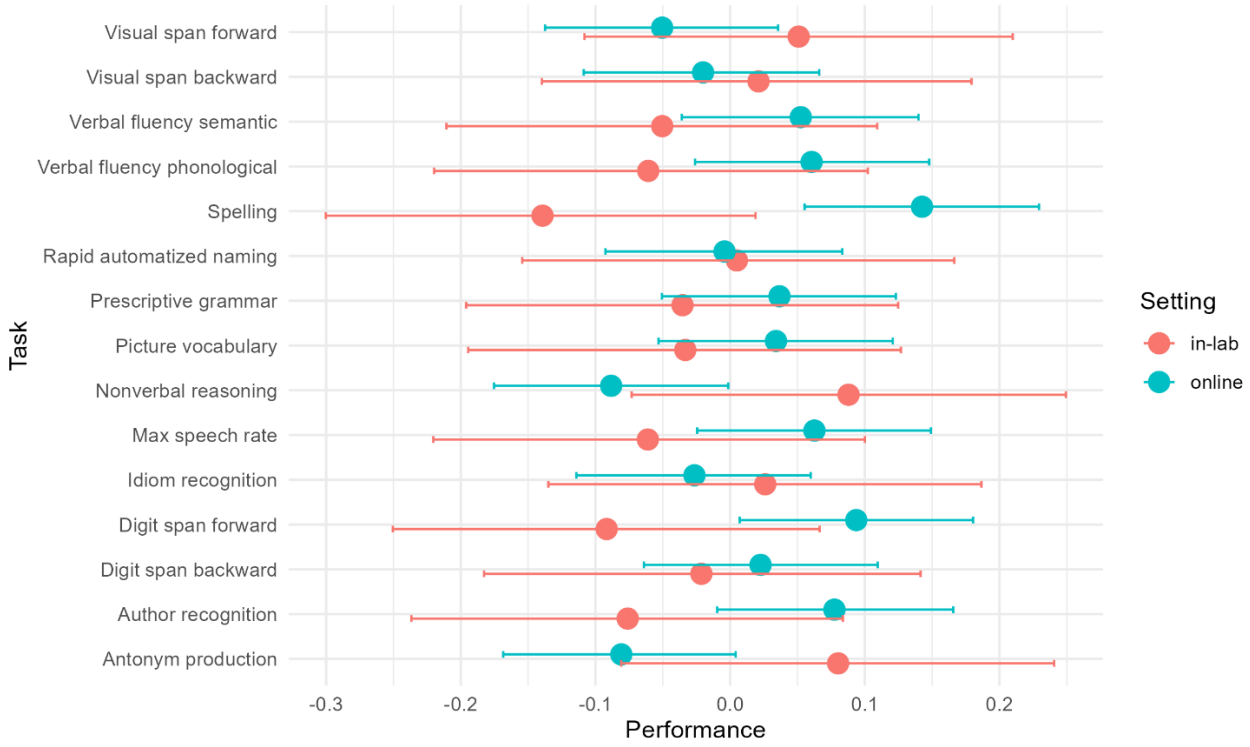

Figure 7: Conditional effects of the interaction between Setting and Task, based on the model with "moderate" priors, reported in Table 5. The points show the model estimate for the given combination of task and setting and the bars visualize the credible intervals. Note that the dependent variable is a task-dependent z-score; the value 0 is relative to the mean of the particular task.

4. Discussion

Collecting data online is increasingly popular in psycholinguistics and psychology in general. Moving data collection online can facilitate experiments with groups that are either geographically dispersed or otherwise less likely to come into a research lab or university. Participants can join the experiment on their own time and data can be collected in parallel without the need for direct supervision of multiple testing sessions. The ease and pragmatism that comes with this type of data collection is accompanied by promising results. Data collected online from many of the most common psychological and linguistic paradigms confirm that results from participants tested online are comparable to those from in-lab samples (Crump et al., 2013; Germine et al., 2012; Hilbig, 2016; Miller et al., 2018, 2018; Patterson & Nicklin, 2023; Semmelmann & Weigelt, 2017). The advantages of online testing are particularly relevant for individual differences studies, which are growing in popularity (Engelhardt et al., 2017; Isbilen et al., 2022; Kidd et al., 2018, 2023; McConnell, 2023; McConnell & Blumenthal-Dramé, 2021; Payne et al., 2014). But importantly, individual differences

paradigms rely on stable estimation of participant ability and thus must detect what group-level designs can safely level out, and it is unclear if research collected online can meet this standard (Hedge et al., 2018).

The Individual Differences in Language Skills (IDLaS-NL) dataset is unique in that it tests samples from the same population (young Dutch-speaking adults under the age of 30, largely in the university context) in the lab and online (Hintz et al., 2020, 2023). It also contains multiple tasks from different domains of linguistic and domain-general skill. Assessing the same population over diverse tasks is perfect for exploring the effect of online testing. We take the subset of this dataset that was collected on the same software for participants in the lab and online, which consists of fifteen tests from three domains: linguistic experience, timed linguistic (production) tasks and untimed linguistic tasks.

We first looked at the descriptive differences between the two settings by comparing the means, medians and standard deviations, as well as the minimum and maximum scores for each task in each condition. These look exceptionally similar, in numeric form as well as in the visualization of the distributions. To confirm that these patterns hold when controlling for participant-based variation, we also modeled the data using a Bayesian hierarchical model. The model shows largely the same patterns that we see in the descriptive statistics; the credible intervals for all interactions between task and setting cross 0 (i.e., the lower bound of the credible interval is below 0 but the upper bound is above 0). Therefore, the model cannot confidently estimate an effect for setting for any task.

There are four tasks, however, which have slightly larger estimates either in the lab or online. One suggests slightly better scores online: the spelling test. For this, participants were shown 60 words (of which half were spelled incorrectly) and had to indicate which were spelled incorrectly. Online, participants scored an average of 5% better (59% vs. 63%, respectively). The maximum recorded score online was 100% correct, whereas in the lab it was 93% correct. The model also predicts higher scores on the spelling task in the online setting, at the rate of approximately one third of the standard deviation on the task (.28 difference between the two settings). While this result may be due to chance (in that better spellers participated in the online setting), it could also be the case that some online participants looked up the spelling of some of words, e.g. on their phones. This would be in line with previous research, which indicates that participants tested online sometimes seek external help in answering challenging questions (Goodman et al., 2013). Importantly, this effect is still minor, although participants were not explicitly discouraged from looking up words.

At the same time, there is little effect of test administration on similar assessments, such as the idiom recognition task. Participants could have just as easily looked up the meaning of an idiom, if

they were striving to maximize their performance. Thus, we cannot conclude that participants looked answers up whenever possible. Perhaps they only sought external help when it was particularly convenient to do so, or when they felt their performance might be judged. For example, being a "bad speller" might be particularly stigmatized compared to not knowing idioms, so participants may have felt compelled to maximize their score on the spelling test only.

Similarly, the digit span task also shows better performance online, where the median is a span of 8 in the lab and 9 online. The model also predicts a difference of .18. Here, participants may have used an aid as simple as a pen and paper to jot down numbers. If this is the case, though, it is interesting that the backward digit span doesn't show a similar trend; the scores on this task do not differ much between settings. This is despite the fact that the backward digit span directly followed the forward digit span. Thus, we cannot say if participants were seeking external aid in this task or if this is simply a spurious result.

On the other hand, two tasks show a trend towards better performance in the lab: antonym production and nonverbal reasoning. Nonverbal reasoning was assessed with Raven's Advanced Progressive Matrices, known to be a challenging task. Participants must look at an example image, try to deduce the pattern they are seeing, then select the correct continuation of the pattern from eight potential items. We know from previous research that difficult tasks lead to higher dropout rates in online tests (Dandurand et al., 2008). In the current study, dropping out would have reduced participants' monetary compensation, which likely encouraged them to complete all tests. Thus, they may have underperformed upon losing motivation for the (twenty-minute long) task. In this task, participants in the lab perform about 3% better in terms of mean scores (60% and 57%, respectively). This is also reflected in the model, with a difference of .17 between the two settings.

We see a similar trend in the antonym production task, though the reason behind this is less clear. Here, the mean scores are similar (80% correct in the lab and 78% correct online), as are the median and maximum scores. The minimum scores, however, differ by nearly 20% (52% in the lab and 36% online). This means that at least one participant online only correctly named the antonym to a third of the trials. Perhaps this task was also perceived as difficult or annoying by some participants. Taken together with performance on the nonverbal reasoning task, these results play into recent discussions of motivation in psycholinguistic research, which suggest that a portion of data is collected from participants who are bored and disengaged (Christianson et al., 2022). In individual differences research, we often operate under the assumption that participants are at peak performance, showcasing their full abilities. Disengagement with the task could be self-selecting to participants with certain personality characteristics, and thus there could be an extraneous variable

in play. However, our results suggest that the impact of this factor is minor and restricted to only some tasks.

Overall, the data hints at two major issues that are surely relevant to individual differences in an online lab: the ability to look up answers online and the likely impact of motivation in a harder task. These two forces have opposite effects: one improves performance online and the other worsens it. However, we have to keep in mind that these differences are small, and the model cannot reliably distinguish them from 0, despite the large sample size. The majority of tasks show no remarkable effect in either direction. The performance online and in the lab is similar, both descriptively and in the model output.

For future research, we recommend that participant motivation not be overlooked. One way of keeping participants motivated, even for harder tasks, might be to "gamify" the tasks, by making them visually appealing and offering symbolic rewards for peak performance (Christianson et al., 2022; Pedersen et al., 2023). We also recommend paying close attention to the clarity and wording of the instructions, in particular asking participants not to look up answers and letting them know that they are not expected to know the correct answer to every question.

Importantly, differences suggesting lower motivation in the online group and the use of external help were only found for a small handful of tasks. We assessed fifteen individual differences assessments from the IDLaS-NL dataset, and eleven of them showed very similar results both in the lab and online. These tasks are diverse; they assess processing speed, visual and auditory working memory, word production, vocabulary and knowledge of prescriptive grammar rules. On all these tasks, there is neither a detectable difference in the model nor notable deviations in the descriptive statistics. Thus, we can confidently support online individual differences research into the domains of word production, domain-general skills, and linguistic experience. There seems good enough reason to suggest venturing forth into the online lab, even for individual differences designs that require precise estimation at the individual level.

4. Conclusion

The current study shows that in a large-scale individual difference battery, performance online and in the lab was nearly indistinguishable. This holds over multiple tasks of different linguistic and psychological constructs, including timed and untimed linguistic tasks and domain-general skills. Although experimenters should keep an eye on both a participant's ability to gain external support

for a task (e.g. through looking up answers) and the likelihood that they will lose motivation, online experimentation seems to be reliable and precise enough to support individual difference research.

Open practices statement: The data and materials for all experiments are available at https://osf.io/2knzc/. The experiment was not was preregistered.

References

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407–1425. https://doi.org/10.3758/s13428-020-01501-5

Araújo, S., Huettig, F., & Meyer, A. S. (2021). What Underlies the Deficit in Rapid Automatized Naming (RAN) in Adults with Dyslexia? Evidence from Eye Movements. *Scientific Studies of Reading*, *25*(6), 534–549. https://doi.org/10.1080/10888438.2020.1867863

Berch, D. B., Krikorian, R., & Huha, E. M. (1998). The Corsi Block-Tapping Task: Methodological and Theoretical Considerations. *Brain and Cognition*, *38*(3), 317–338. https://doi.org/10.1006/brcg.1998.1039

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Blake, A., & Dąbrowska, E. (2024). Investigating the relationship between the speed of automatization and linguistic abilities: Data collection during the COVID-19 pandemic. *Linguistics Vanguard*, *0*(0). https://doi.org/10.1515/lingvan-2021-0145

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Brysbaert, M., Sui, L., Dirix, N., & Hintz, F. (2020). Dutch Author Recognition Test. *Journal of Cognition*, *3*(1), 6. https://doi.org/10.5334/joc.95

Bürkner, P.-C. (2017). **brms**: An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. https://doi.org/10.3758/s13428-013-0365-7

Christianson, K., Dempsey, J., Tsiola, A., & Goldshtein, M. (2022). What if they're just not that into you (or your experiment)? On motivation and psycholinguistics. In *Psychology of Learning and Motivation* (Vol. 76, pp. 51–88). Elsevier. https://doi.org/10.1016/bs.plm.2022.03.002

Corps, R. E., & Meyer, A. S. (2023). Word frequency has similar effects in picture naming and gender decision: A failure to replicate Jescheniak and Levelt (1994). *Acta Psychologica*, *241*, 104073. https://doi.org/10.1016/j.actpsy.2023.104073

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, *2*(3), 219–253. https://doi.org/10.1075/lab.2.3.01dab

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434. https://doi.org/10.3758/BRM.40.2.428

Denckla, M. B., & Rudel, R. G. (1976). Rapid 'automatized' naming (R.A.N.): Dyslexia differentiated from other learning disabilities. *Neuropsychologia*, *14*(4), 471–479. https://doi.org/10.1016/0028-3932(76)90075-0

Dunn, L. M., & Dunn, D. (1997). *Peabody Picture Vocabulary Test (3rd Edition)*. American Guidance Service.

Engelhardt, P. E., Nigg, J. T., & Ferreira, F. (2017). Executive Function and Intelligence in the Resolution of Temporary Syntactic Ambiguity: An Individual Differences Investigation. *Quarterly Journal of Experimental Psychology*, *70*(7), 1263–1281. https://doi.org/10.1080/17470218.2016.1178785

Fairs, A., & Strijkers, K. (2021). Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors. *PLOS ONE*, *16*(10), e0258908. https://doi.org/10.1371/journal.pone.0258908

Garcia, R., Roeser, J., & Kidd, E. (2022). Online data collection to address language sampling bias: Lessons from the COVID-19 pandemic. *Linguistics Vanguard*, *0*(0). https://doi.org/10.1515/lingvan-2021-0040

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. https://doi.org/10.1002/bdm.1753

Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behavior Research Methods*, *52*(6), 2283–2286. https://doi.org/10.3758/s13428-020-01395-3

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. (2020). *Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/xr7y3

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

He, J., Meyer, A. S., Creemers, A., & Brehm, L. (2021). Conducting Language Production Research Online: A Web-based Study of Semantic Context and Name Agreement Effects in Multi-Word Production. *Collabra: Psychology*, *7*(1), 29935. https://doi.org/10.1525/collabra.29935

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hilbig, B. E. (2016). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. https://doi.org/10.3758/s13428-015-0678-9

Hintz, F., Dijkhuis, M., van 't Hoff, V., Huijsmans, M., Kievit, R. A., McQueen, J., & Meyer, A. S. ((in preparation)). *Assessing the principal dimensions of speaking and listening skills*.

Hintz, F., Dijkhuis, M., van 't Hoff, V., McQueen, J. M., & Meyer, A. S. (2020). A behavioural dataset for studying individual differences in language skills. *Scientific Data*, *7*(1), 429. https://doi.org/10.1038/s41597-020-00758-x

Hintz, F., Shkaravska, O., Dijkhuis, M., Van 'T Hoff, V., Huijsmans, M., Van Dongen, R. C. A., Voeteé, L. A. B., Trilsbeek, P., McQueen, J. M., & Meyer, A. S. (2023). IDLaS-NL – A platform for running customized studies on individual differences in Dutch language skills via the Internet. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02156-8

Hubers, F., Cucchiarini, C., Strik, H., & Dijkstra, T. (2019). Normative Data of Dutch Idiomatic Expressions: Subjective Judgments You Can Bank on. *Frontiers in Psychology*, *10*, 1075. https://doi.org/10.3389/fpsyg.2019.01075

Isbilen, E. S., McCauley, S. M., & Christiansen, M. H. (2022). Individual differences in artificial and natural language statistical learning. *Cognition*, *225*, 105123. https://doi.org/10.1016/j.cognition.2022.105123

Kidd, E., Arciuli, J., Christiansen, M. H., & Smithson, M. (2023). The sources and consequences of individual differences in statistical learning for language development. *Cognitive Development*, *66*, 101335. https://doi.org/10.1016/j.cogdev.2023.101335

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences*, *22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Mainz, N., Shao, Z., Brysbaert, M., & Meyer, A. S. (2017). Vocabulary Knowledge Predicts Lexical Processing: Evidence from a Group of Participants with Diverse Educational Backgrounds. *Frontiers in Psychology*, *8*, 1164. https://doi.org/10.3389/fpsyg.2017.01164

McConnell, K. (2023). Individual Differences in Holistic and Compositional Language Processing. *Journal of Cognition*, *6*(1), 29. https://doi.org/10.5334/joc.283

McConnell, K., & Blumenthal-Dramé, A. (2021). Usage-Based Individual Differences in the Probabilistic Processing of Multi-Word Sequences. *Frontiers in Communication*, *6*, 703351. https://doi.org/10.3389/fcomm.2021.703351

Miller, R., Schmidt, K., Kirschbaum, C., & Enge, S. (2018). Comparability, stability, and reliability of internet-based mental chronometry in domestic and laboratory settings. *Behavior Research Methods*, *50*(4), 1345–1358. https://doi.org/10.3758/s13428-018-1036-5

Patterson, A. S., & Nicklin, C. (2023). L2 self-paced reading data collection across three contexts: In-person, online, and crowdsourcing. *Research Methods in Applied Linguistics*, *2*(1), 100045. https://doi.org/10.1016/j.rmal.2023.100045

Payne, B. R., Grison, S., Gao, X., Christianson, K., Morrow, D. G., & Stine-Morrow, E. A. L. (2014). Aging and individual differences in binding during sentence understanding: Evidence from temporary and global syntactic attachment ambiguities. *Cognition*, *130*(2), 157–173. https://doi.org/10.1016/j.cognition.2013.10.005

Pedersen, M. K., Díaz, C. M. C., Wang, Q. J., Alba-Marrugo, M. A., Amidi, A., Basaiawmoit, R. V., Bergenholtz, C., Christiansen, M. H., Gajdacz, M., Hertwig, R., Ishkhanyan, B., Klyver, K.,

Ladegaard, N., Mathiasen, K., Parsons, C., Rafner, J., Villadsen, A. R., Wallentin, M., Zana, B., & Sherson, J. F. (2023). Measuring Cognitive Abilities in the Wild: Validating a Population-Scale Game-Based Cognitive Assessment. *Cognitive Science*, *47*(6), e13308. https://doi.org/10.1111/cogs.13308

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Pronk, T., Hirst, R. J., Wiers, R. W., & Murre, J. M. J. (2022). Can we measure individual differences in cognitive measures reliably via smartphones? A comparison of the flanker effect across device types and samples. *Behavior Research Methods*, *55*(4), 1641–1652. https://doi.org/10.3758/s13428-022-01885-6

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven Manual: Section 4, Advanced Progressive Matrices*. Oxford Psychologists Press Ltd.

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*(2), 309–327. https://doi.org/10.3758/s13428-014-0471-1

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, *134*, 104472. https://doi.org/10.1016/j.jml.2023.104472

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, Hosting and Recruiting: A Brief Introduction to Running Behavioral Experiments Online. *Brain Sciences*, *10*(4), 251. https://doi.org/10.3390/brainsci10040251

Schlichting, L. (2005). *Peabody Picture Vocabulary Test Dutch-III-NL.* Harcourt Assessment BV.

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*(4), 1241–1260. https://doi.org/10.3758/s13428-016-0783-4

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00772

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods*, *46*(1), 95–111. https://doi.org/10.3758/s13428-013-0345-y

Wechsler, D. (2004). *Wechsler Adult Intelligence Scale, 3rd Edition (WAIS-III)*.