

## **Worldlikeness: A Web application for typological psycholinguistics**

Tsung-Ying Chen\*

Department of Foreign Languages and Literature, National Tsing Hua University, Taiwan

No. 101, Sec. 2, Guangfu Rd., Hsinchu, Taiwan, R.O.C. 300

chen.ty@mx.nthu.edu.tw +886-5-5715131#62025

James Myers

Graduate Institute of Linguistics, National Chung Cheng University, Taiwan

**Abstract** Worldlikeness, currently hosted online at <https://worldlikeness.org>, is a free, open-source Web-based experimental tool and database that seeks to make it easier to extend the scope of cross-linguistic studies on language processing. Unlike the traditional approach to typological psycholinguistics that requires a large, coordinated team of language experts and psycholinguists, Worldlikeness allows individual psycholinguists to collect data on individual languages and then share them to help build a cross-linguistic database. As the database grows, outside researchers can easily conduct their own typological cross-linguistic analyses and contribute to our understanding of language-specific and universal properties of language processing. In this paper, we first introduce the key features of Worldlikeness, including stimulus display, reaction time measurement, and privacy protection, and then demonstrate the reliability of Worldlikeness in its default task: wordlikeness judgments (judgments of nonce word acceptability).

**Keywords** Typology, Psycholinguistics, Wordlikeness, Megastudy, Web experiment

Languages have many subtle factors that are not governed by universal principles and thus lead to language-specific learning experiences, which in turn drive idiosyncratic language processing (e.g., speech perception: Guion et al., 2000; Strange, 1995; morphological processing: Jarema et al., 1999; Zhou & Marslen-Wilson, 2000; sentence processing: Bates 1999; Jaeger & Norcliffe, 2009; Norcliffe et al., 2015; Wells et al., 2009). A primary goal in psycholinguistic studies is therefore to understand which factors are more important than others in different languages and how they could interact with universal principles to account for the processing performance observed in speakers of individual languages. Taking phonological processing as an example, it has been widely acknowledged that while all languages presumably have separate levels for phoneme, mora, and syllable (e.g., Hyman, 1985; McCarthy & Prince, 1986), native speakers of different languages may not weight information at these different levels equally in their production and perception: Mandarin and Cantonese speakers treat the syllable as most important (e.g., Chen et al., 2002; O'Seaghdha et al., 2010), Japanese speakers rely on mora counting in many phonological contexts more intensively (e.g., Han, 1994; Otake et al., 1993), while English speakers tend to be affected more by phoneme-level factors like transition probabilities (e.g., Frisch et al., 2000, Hayes & White, 2013; O'Seaghdha et al., 2010).

The root cause of such processing differences, however, cannot be confirmed easily due to multiply confounded variables in any small sample group of languages. For example, English and Mandarin differ not just in the number of phonemes (high in English, low in Mandarin), but also in the number of syllables (high in English, low in Mandarin) and orthography (phoneme-based letters vs. syllable-based logographs), any of which may be responsible for differences in how the two languages treat phonemes and syllables. Given that the rule of thumb in regression analysis is to have at least ten data points per independent variable (e.g., Harrell, 2015, p. 72), typological

psycholinguistics requires far more than small convenience samples of languages. This would extend the megastudy philosophy from lexical processing within a single language (Balota et al., 2012; Keuleers & Balota, 2015) to allow for languages themselves, not just lexical items, to be treated as a random variable, in what Myers (2016) dubs meta-megastudies.

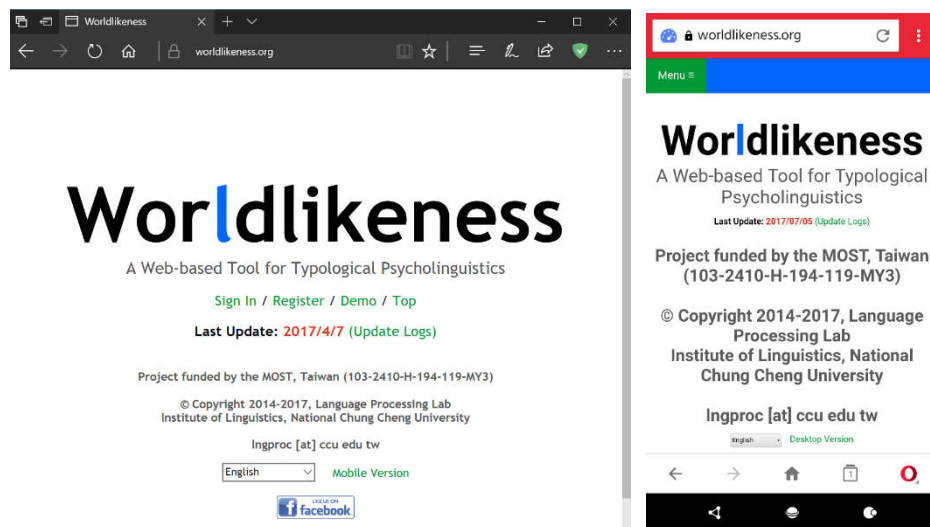
However, quantitatively sophisticated typological psycholinguistics poses serious practical challenges, requiring collaboration between research teams and language experts in different locations to collect data in traditional lab settings, all using the same experimental design. For instance, the study by Bates et al. (2003) on factors affecting picture naming in just seven languages (Bulgarian, English, German, Hungarian, Italian, Mandarin, and Spanish) involved 22 authors, and even with this great effort, the sampled languages still cover a very small range of typological variation, making it impossible to disentangle partially confounded cross-linguistic variables, including those mentioned above (syllables and orthography).

### **Typological psycholinguistics with web crowdsourcing and data sharing**

In this paper, we describe an online tool intended to streamline collaboration so that independent psycholinguists can conduct research on individual languages, but do so in a system that enforces methodological consistency and encourages data sharing with outside typological researchers. In other words, just as with typological research on grammar, which involves compilations of grammatical descriptions originally created by independent researchers (e.g., Haspelmath et al., 2005), we suggest that the scope of typological psycholinguistics would be greatly expanded by distinguishing data collection from cross-linguistic meta-analysis. The key for typological psycholinguistics is thus to encourage more people to run consistent studies and to share their data,

not to create ever-larger research teams (cf. the ManyBabies project on child language within the Web-based Open Science Framework, a Web-based system for enabling collaboration among dozens to hundreds hundreds of researchers; Frank et al., 2017: <https://osf.io/rpw6d/>).

Worldlikeness (Fig. 1) is a free web application implementing this concept. It is aimed at three types of users: experimenters, who can run experiments online or upload the results of old experiments, their participants, and researchers, who use the shared experimental data to run cross-linguistic analyses. The application is designed with a focus on the factors affecting wordlikeness, or the intuitive acceptability of non-words by native speakers (hence the punning name of the web application), but it can be easily extended to investigate different aspects of human language processing using any experimental paradigms with a simple stimulus-response trial structure (e.g., non-primed lexical decision or perceptual discrimination). As we continue to motivate experimenters and participants to share their data via Worldlikeness, the Web platform is expected to grow into a large online database of psycholinguistic judgments and response latencies.



**Fig. 1** Screenshots of the Worldlikeness home page in the Microsoft Edge desktop browser (left) and the Opera Mini mobile browser (right)

There are a growing number of Web-based tools allowing behavioral researchers to run experiments online (e.g. ttool: von Bastian et al., 2013; turktools: Erlewin & Kotek, 2016; YourMorals.Org: Graham et al., 2011; WebExp: Keller et al., 2009; Amazon Mechanical Turk: Paolacci et al., 2010; PsychoJS: Peirce, 2009; jsPsych: de Leeuw, 2015; TaskPrime.com: Litman et al., 2016). The increase in convenience and data size does not seem to be associated with a severe decrease in data quality (e.g. Buhrmester et al., 2011; Crump et al., 2013; Goodman et al., 2013; Goslin et al., 2004). Worldlikeness supplements these tools with the following distinctive features, designed to increase ease of use, ethical research practices, and data sharing (further technical details will be spelled out later in the paper).

*Accessibility* – Worldlikeness is a ready-to-use Web application developed using the JavaScript-based programming language Meteor (Coleman & Greif, 2016) integrated with the server-side database package MongoDB (MongoDB, Inc., 2008-2017), which can be accessed in major modern desktop and mobile device web browsers (e.g., Microsoft Internet Explorer, Mozilla Firefox, and Google Chrome). Experimenters need no additional plug-ins, programming, or server management skills to create, manage, and run their Web experiments in Worldlikeness. Worldlikeness allows experimenters to invite and run participants in any language (i.e., in the consent form and instructions, in addition to the experimental stimuli themselves). The main interface itself is currently available in English and Mandarin, with other languages to be added over the next few years.

*Ethics* – In the Worldlikeness project, we have put great effort into developing a system that conforms to the core ethical and application principles in the Belmont report (Ryan et al., 1979): respects for persons, beneficence, and justice. Thus, all users have complete control over their own data, both personal and experimentally elicited.

Experimenters only need to supply an e-mail address to create an account, and participants and outside researchers remain completely anonymous while participating in an experiment or downloading publicly shared data sets. In addition, Worldlikeness helps experimenters prepare their experiment description, consent form, and instructions, and rewards their participants with a report summarizing their experimental results. Experimenters have full control of whether to openly or privately recruit participants and whether to share their results. Participants ultimately decide whether to authorize their experimental data to users other than the original experimenter, and they can remove their data at any time during or after an experimental session.

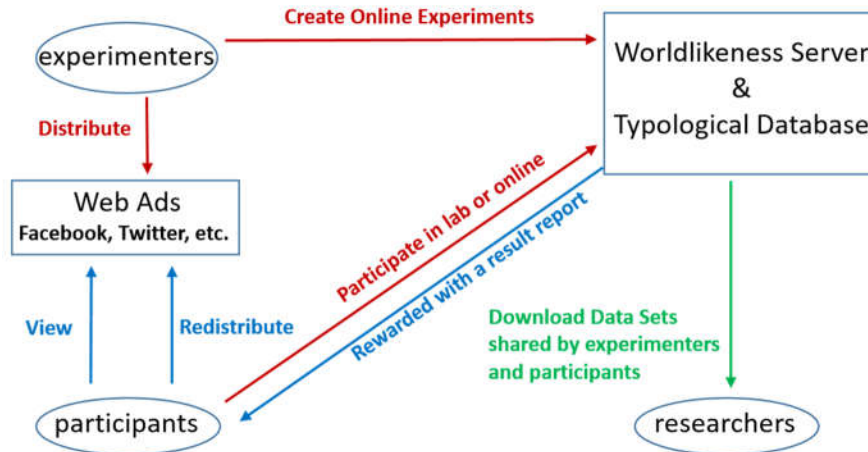
*Functionality* – Worldlikeness allows experimenters to incorporate and adjust elements commonly included in behavioral experiments, including the use of an eye fixation mark and its display duration, trial duration, stimulus duration, stimulus type (text, audio, image, video), stimulus size, and response type (binary vs. seven-point Likert-scale judgment). Experimenters can either force participants to respond by pressing a key on the keyboard, or allow them to click buttons on the screen with a mouse or via a touchscreen. Experimenters are also able to insert a practice session prior to the formal session. Worldlikeness automatically records information crucial to an analysis of participants' performance, including reaction times, key responses, browser type, and the time at the beginning and the end of an experimental session.

*Consistency* – In Worldlikeness, all experiments have a similar judgment task design to minimize variation across studies and thus reduce confounds with cross-linguistic variation in typological analyses (for artifacts due to task differences, see, e.g., Gerrits & Schouten, 2004).

In the rest of this paper, we will elaborate on the above features in greater detail, and report evidence concerning the reliability of data collected via Worldlikeness.

### **The basic concept of the Worldlikeness ecosystem**

Worldlikeness is a Web experiment platform as well as a typological psycholinguistics database developed around three different user roles, as illustrated in Fig. 2. *Experimenters* create and design their web experiments in the Worldlikeness online server, or upload and share experimental results from their previous studies. *Participants* are speakers/signers of a target language recruited by experiments, for example via a Web advertisement distributed across internet communities. They may participate in a Worldlikeness experiment either in a traditional lab setting or anonymously via their own device connected to the internet. Motivated participants (see below) are expected to help redistribute the web experiments and speed up the Web crowdsourcing process. *Researchers* are ‘meta-analyzers’ interested in studying cross-linguistic language processing, who can download experimental data sets from multiple languages that have been publicly shared in Worldlikeness.



**Fig. 2** The Worldlikeness ecosystem

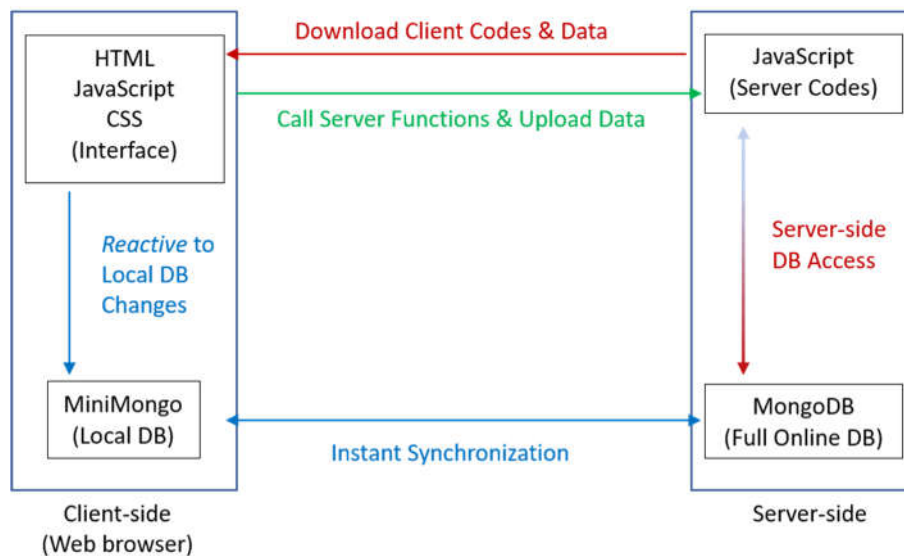
In order to foster continuous growth in this Web experiment ecosystem, Worldlikeness provides incentives for each user role. Experimenters are motivated to share their experimental data in exchange for a larger quota to run new experiments. Participants are guided to authorize their experimental data to be accessible to all users of Worldlikeness and are also encouraged to share experiment announcements with friends via the internet. Researchers can download more publicly shared experimental data by signing up for an experimenter account, which may encourage them to run experiments on their own languages. We will explain the design and efficacy of these incentives after first describing the technical specifications of the system.

### **The Meteor framework**

Meteor is a Web development suite encapsulating a Web server and a server-side MongoDB database; the basic framework is illustrated in Fig. 3. A Web application created and hosted using Meteor contains both client codes and server codes. The client codes include (i) HTML and CSS files to present the application's user interface in modern Web browsers, and (ii) JavaScript files



with server communication functions and user interface functions. No additional client-side plug-in or software is required to use a Meteor application. The server JavaScript codes handle the communication between the client and the online database. When users are connected to a Meteor application, the client-side files will first be loaded into the users' browser, and the application also creates a local cache of the online MongoDB database in the browser, which is called MiniMongo. MiniMongo and its online master database are synchronized automatically upon any changes made to either of them, and the application's user interface is *reactive* to the changes in MiniMongo. That is, a Web page accessing information from the database is updated automatically when the information is changed, and a client-side Meteor function is re-executed on its own as well if the function retrieves the information from the database.



**Fig. 3** The Meteor framework

Under this framework, Worldlikeness only downloads client files for its user interface and caches a small amount of the system and user data in the browser upon users' initial access to the

Web application. In particular, Worldlikeness only sends 1.18 megabytes of data to the users' browser to load its front page as estimated by Pingdom Website Speed Test (test run on Oct 20, 2017 at <https://tools.pingdom.com>). Large data sets, such as experimental materials and publicly shared data sets, are downloaded only upon request, and we continue to put effort into minimizing data transmission between users and the Worldlikeness server to make the application more accessible. This is especially important for participants speaking an understudied language in areas with a less stable internet connection (see the 'Mobile Interface Development' section below).

### **The Worldlikeness application design**

*Online security and research ethics* – Online privacy and data security have been our primary concerns during the development of Worldlikeness. Worldlikeness requests minimal personal information from the three types of users. Aside from their e-mail address, experimenters are not asked to provide any additional personal information, and the e-mail address itself is only used to verify their experimenter account, but is never exposed within or beyond Worldlikeness. The password is hashed and encrypted for every experimenter account in the Worldlikeness server. Participants and researchers do not have to create an account in Worldlikeness at all to participate in an experiment or to download shared data, and thus remain completely anonymous throughout their interactions with the system. Worldlikeness does record the IP address of all users, which browsers automatically make public to Web sites anyway, but this is used primarily to allow experimenters to examine whether participants seem to be attempting to re-take the same experiment. The Worldlikeness administrators (currently, just the authors of this paper) are also capable of retrieving all users' IP address to monitor and possibly block abnormal user behavior.

All visits to Worldlikeness are connected via the HTTPS protocol with an SSL certificate issued by Let's Encrypt (<https://letsencrypt.org>) to significantly reduce the risk of accidentally leaking personal and experimental data to a third party.

The partial/full anonymity in Worldlikeness is part of the project's commitment to follow standard research ethics guidelines, which require delinking participants' data from their identity. In addition, Worldlikeness gives its users full control over how to store and authorize their data online. Worldlikeness features three data authorization options on the online consent form page (Fig. 4), whereby participants can choose to authorize access to their experimental results to (i) all Worldlikeness users (including guest users like typological researchers), (ii) registered users only (i.e. those with a verified experimenter account), or (iii) the corresponding experimenters only (i.e., those running the experiment).

Statement of Consent: I have read the above information, and have received answers to any questions I asked from the experimenters in person or by emails. I understand that I will remain TOTALLY ANONYMOUS and my personal information and experiment data will be stored on the Worldlikeness server and accessed by user groups of my choice below:

All Users  Registered Users Only  This Experimenter (Research Team) Only

By clicking the "Agree" button below, I am agreeing to take part in the study. Clicking the "Cancel" button will take you back to the home page.

Agree Cancel

**Fig. 4** Data authorization options on the consent form page in the Worldlikeness desktop interface

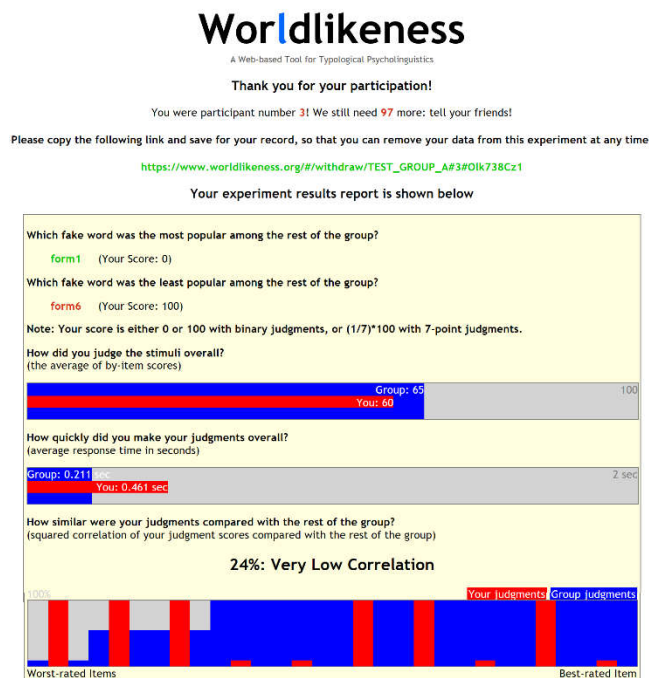
Once the decision is made, it is stored in the Worldlikeness server to be checked in the server-side data-retrieval Meteor function, which cannot be overridden under any circumstances. Anonymous background information is only collected after participants choose their authorization option and give their informed consent. Participants also receive a link after they complete an experimental session, which they can use to withdraw their data from any experiment at any time,

without needing prior approval from the experimenter or the Worldlikeness administrators (Fig. 5).



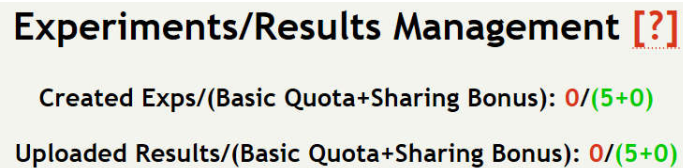
**Fig. 5** Withdrawing data via a direct link

Worldlikeness also help participants understand their rights and responsibilities before participating in an experiment by requiring experimenters to provide a detailed description of the experiment and a consent form. It is also mandatory for experimenters to provide a link to their public research webpage in the description, so participants will have a chance to send inquiries to experimenters. After participants complete an experimental session, they are rewarded with a report that summarizes their own performance (Fig. 6). This report includes the participant's most and least favored items, a comparison of their mean reaction times and judgment scores with other participants in the experiment, and Pearson's coefficient of determination ( $r^2$ ) comparing the participant's by-item judgment scores with the experimental group's by-item mean judgment scores. This report contributes to the online participant recruitment process by "gamifying" the experiment in the manner of social media, encouraging participants to distribute the Web advertisement for the experiment among their friends so that they can all compare their results (see the section 'Wordlikeness judgment of Mandarin-Southern Min bilinguals' below).



**Fig. 6** Results report for participants at the end of an experimental session

*Incentives for sharing data and running more experiments* – As mentioned earlier, in order to expand the online typological psycholinguistics database, the ecosystem of Worldlikeness must be supported with continuous growth in the number of publicly shared data sets as well as experiments of different languages. Worldlikeness is thus designed to motivate its users both explicitly and implicitly to help in this goal. For experimenters, the Worldlikeness policy can be summarized in the motto: ‘the more you share, the more you can do’. Every experimenter account is assigned a basic quota that allows them to create five experiments, plus upload five experimental data sets from their previous studies (Fig 7). The experimenters’ quota will increase by one for every data set collected via, or uploaded to, Worldlikeness, that is made public. This sharing bonus is made explicit in the dashboard on the experiments/results management page (see Fig. 7). Currently there is still a maximum quota, but it is quite high: 25 Worldlikeness-run experiments plus 25 uploaded data sets.



**Fig. 7** Experiments and results dashboard

For participants, the user interface is designed to implicitly encourage them to share their experimental data as well. First, the most open authorization option (i.e. allowing access to all Worldlikeness users) is always selected by default on the consent form page (see Fig. 4 above), which is intended to bias participants against the more conservative authorization options (e.g., Johnson et al., 2002; Park et al., 2000; cf., Löfgren et al., 2012). Second, we experimentally tested the optimal order for the three authorization options in the desktop interface so that participants would be least likely to switch from the default and most open option to the other two more conservative options. In our wordlikeness judgment experiments (see the ‘Wordlikeness judgment by Mandarin-Southern Min bilinguals’ section below for the experimental results), we recruited 81 in-lab participants and 156 online participants, who were randomly assigned to two choice order conditions. In the first condition, the conservativeness of data authorization decreased from left to right (the direction in which horizontal Chinese text is generally written) as shown previously in Fig. 4 (the ‘Open First’ condition). The alternative test layout was the exact opposite, with the most open (default) option on the right and the most conservative option on the left (the ‘Open Last’ condition), with the intermediate option in the middle. In both conditions, the most open option was selected by default, and the participants made their own decisions regarding data authorization upon signing the online consent form in the desktop interface without any further interference.

For the in-lab participants in Table 1, most did not switch from the default (most open) option (i.e. all users) to either of the other two less open options (i.e. only registered users or only corresponding experimenters) (one-way Pearson’s chi-squared test:  $\chi^2(1) = 18.78, p < .001$ ). This pattern contrasts sharply with participants’ choices in a previous version of Worldlikeness described in Authors (2017), in which the default was the most conservative option (authorizing the data only to corresponding experimenters), possibly leading most participants to stick with this option. The new default option design should therefore help make data more available for typological analyses.

This bias to keep the default open option was enhanced if this option appeared in the leftmost position of the sequence (two-way Pearson’s chi-squared test with Yates’ continuity correction:  $\chi^2(1) = 8.1, p < .01$ ), consistent with previous studies showing primacy effects in choice, that is, participants’ preference for the first option in a sequence of choices (e.g. Miller & Krosnick 1998; Mantonakis et al., 2009). Thus, in the current desktop version of Worldlikeness, the most open authorization option appears on the left in both English and Chinese interfaces. This default open/first strategy will be maintained in when we add interfaces for languages written from right to left (e.g., Arabic and Hebrew).

	Most open	Less open	Total
Open First	35	4	39
Open Last	25	17	42
Total	60	22	81

Table 1. Distribution of in-lab participants’ authorization choices by the order of options

The effect of selecting the most option by default was also found for online participants (Table 2) (one-way Pearson’s chi-squared test:  $\chi^2(1) = 12.41, p < .001$ ). However, the primacy

effect did not interact with the default effect (two-way Pearson’s chi-squared test with Yates’ continuity correction:  $\chi^2(1) = 0.51, p = .47$ ). Our speculation is that we explained the three options to the in-lab participants more carefully, therefore enhancing the primacy effect by accident, whereas no additional explanation was given to the online participants on the consent form page. Nevertheless, there was still a non-significant trend in the online participants, with the most open option being favored slightly more when it appeared on the left. We plan to run the same test specifically for the mobile interface in a future test, in which the three authorization options are aligned vertically, with the most open and default option on the top (therefore the first option) and the most conservative one at the bottom.

	Most open	Less open	Total
Open First	52	25	77
Open Last	48	31	79
Total	100	56	156

Table 2. Distribution of online participants’ authorization choices by the order of options

Participants’ choices serve as an incentive for researchers as well, since participants have the option to authorize their data only to registered users, thereby restricting data to outsiders who do not contribute data themselves; Fig. 8 illustrates the dramatic difference in the number of participants available to users of different status. Therefore, if researchers need more data to tease apart the effects of typological variables, they may be motivated to sign up for an experimenter account. This will give them a chance to run new experiments or upload old data, enhancing the Worldlikeness database for other users as well.

Exp Short Title [?]	WL Exp [?]	Download [?]	Participants [?]
閩南語聽覺似詞判斷(PS) [Southern Min]	Yes	<a href="#">Download</a>	20



Exp Short Title [?]	WL Exp [?]	Download [?]	Participants [?]
閩南語聽覺似詞判斷(PS) [Southern Min]	Yes	<a href="#">Download</a>	15
Exp Short Title [?]	WL Exp [?]	Download [?]	Participants [?]
閩南語聽覺似詞判斷(PS) [Southern Min]	Yes	<a href="#">Download</a>	7

**Fig. 8** Different numbers of participants based on users' identity (top = corresponding experimenters, central panel = registered and verified users, bottom = guest users)

*Experiment management and design* – Wordlikeness provides a user-friendly graphic interface to help experimenters manage and design their experiments more easily. When experimenters log into Worldlikeness, they immediately see the record of recent activities involving their experiments (Fig. 9), which gives them a quick overview of their current progress and the popularity of their shared experimental data. The system also allows experimenters to add up to 20 collaborators to an experiment (Fig. 10) to help run and manage the experiment. Collaborators can, but do not have to be, fellow linguists testing different languages, thus allowing for traditional big-team collaboration without requiring it. On the experimental results management page, the main experimenter ('E') is the one who provides participants and researchers with a public Web page containing contact information, a 'guest collaborator' ('G') can only view the experimental settings and download experiment results, and a 'core collaborator' ('C') can also change the experimental settings and run the experiment (Fig. 11).

**Recent Experiment Activities [?]**

11-20 of 68 Entries

Log Type:  Number per Page:  [Next 10](#) [Prev 10](#)

Participation: "Exp "閩南語假字似詞判斷(KY-MIN-ONLINE)" run by guest-#LyB2Yp5PK5A6yDh" @ Fri Jun 16 2017 10:39:34 GMT+0800 (Taipei Standard Time)
Participation: "Exp "DEMO EXPERIMENT" run by guest-#oReeZy7rvo7mkm4f9" @ Fri Jun 16 2017 01:38:56 GMT+0800 (Taipei Standard Time)
Participation: "Exp "閩南語假字似詞判斷(KY-MIN-ONLINE)" run by guest-#D7jK99FkYgAEpf73a" @ Thu Jun 15 2017 21:22:41 GMT+0800 (Taipei Standard Time)
Participation: "Exp "閩南語假字似詞判斷(KY-MIN-ONLINE)" run by guest-#taJ58yv6y6qkSqrBB" @ Thu Jun 15 2017 17:24:08 GMT+0800 (Taipei Standard Time)
Collaboration: "Exp ratings #1 of "圖樣感覺判斷實驗(TWO STROKE)" submitted by lngproc@... @ Thu Jun 15 2017 16:12:37 GMT+0800 (Taipei Standard Time)
Collaboration: "Exp "圖樣感覺判斷實驗(TWO STROKE)" completed by lngproc@... #2" @ Thu Jun 15 2017 16:12:10 GMT+0800 (Taipei Standard Time)

**Fig. 9** The list of experiment logs in Worldlikeness

**Add a collaborative experimenter (email account): [?]**

Core  Guest [Add](#)

[fake@core.colla \(Core\) X](#) [fake@guest.colla \(Guest\) X](#)

**Fig. 10** Adding collaborators on the experimental settings page





<input type="checkbox"/>	CHEWING GUM TEST (TEXT) <b>G</b>	<a href="#">Change</a>	<a href="#">Download</a>	2 / 10
<input type="checkbox"/>	TESTING PRACTICE <b>E</b>	<a href="#">Change</a>	<a href="#">Download</a>	28 / 100
<input type="checkbox"/>	VIDEO/IMAGE <b>C</b>	<a href="#">Change</a>	<a href="#">Download</a>	0 / 100
<input type="checkbox"/>	SOUNDS <b>C</b>	<a href="#">Change</a>	<a href="#">Download</a>	0 / 100

**Fig. 11** Labels for different collaborator roles in the list of experiments

The experimental design components of Worldlikeness follow from its research goals to encourage a regression-based design in the study of lexical processing (e.g., wordlikeness judgments), since lexical variables tend to be confounded and gradient (Baayen, 2010; Cutler, 1981). For example, phonological wordlikeness judgments are influenced by phonological neighborhood density (overall similarity to lexical words) and phonotactic probability (phoneme

transition probability in lexical words), which are noncategorical and correlated (Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997). Such considerations have been a major impetus for the regression-based megastudy movement (Balota et al., 2004; Kuperman, 2015). The regression-based approach also means that researchers are free to try out alternative quantifications for their predictor variables, many of which have been proposed for both phonological neighborhood density and phonotactic probability (see, e.g., Albright, 2009; Bailey & Hahn, 2001; Myers, 2015; Yarkoni et al., 2008). Therefore, Worldlikeness does not expect stimuli to be divided into separate groups according to any categorical factor, although experimenters are still free to include category indications in uploaded stimulus files, and to include them in the analyses of their results. Consistent with this regression-based megastudy movement, Worldlikeness does not limit the number of stimuli in an experiment, and experimenters can choose to present only a small random subset of stimuli to each participant to avoid the “judgment fatigue” (Synder, 2000) that may occur during a lengthy experimental session.

Worldlikeness also supports building experiments with a counterbalanced Latin square design (e.g., in the context of acceptability judgments, to avoid artifacts from cross-trial priming as well as judgment fatigue). To do this, experimenters can add a group label to up to eight sub-experiments (Fig. 12). Worldlikeness will then automatically assign participants to each experiment in the group on a rotating basis. Participants will always be assigned to a sub-experiment with the fewest number of participants, or the first sub-experiment in the list of those with the same fewest number of participants. Each participant is given an ID unique across all sub-experiments. This function also allows for cross-group factorial designs.

中文視覺假字判斷(SQUARE D)  PRIMING	Change	Download	15/50
中文視覺假字判斷(SQUARE C)  PRIMING	Change	Download	15/50
中文視覺假字判斷(SQUARE B)  PRIMING	Change	Download	16/50
中文視覺假字判斷(SQUARE A)  PRIMING	Change	Download	16/50

**Fig. 12** Group labels (‘PRIMING’ in the purple background) in Worldlikeness

For each Worldlikeness experiment, experimenters can use their own device to create a tab-delimited or comma-separated list of stimuli with one stimulus per row and one variable per column, and upload it to Worldlikeness (Table 3). Worldlikeness will then automatically assign a numeric label to each stimulus, and present the stimuli to each participant in a different random order.

The Web interface of Worldlikeness is developed following the HTML5 standard, which makes it possible to use different types of stimuli in unimodal and cross-modal language processing experiments. Worldlikeness thus allows experimenters to use Unicode text, images (JPEG, BMP, and PNG files), sounds (MP3 and WAV files), and videos (MP4 files, e.g., for the study of sign languages) as their stimuli. In their list of stimuli, experimenters can include a textStimuli column for Unicode text, and/or columns with file paths to the multimedia files. Text-sound cross-modal experiments are made possible by including both ‘textStimuli’ and ‘audioStimuliPath’ columns (image-sound pairing will also be made available in future updates). All multimedia files are always loaded to the participants’ browser prior to the beginning of an experimental session, so that the measurement of reaction times is not contaminated by network lags. The duration of multimedia files is also measured automatically in Worldlikeness and can be downloaded by Worldlikeness users as part of the results. By adding the ‘Session’ column in the stimulus list, each stimulus can be specified as part of a practice or formal session, and Worldlikeness can separate the two sessions accordingly. In addition to these parameter columns,

experimenters are allowed to attach other columns as supplementary information (e.g., lexical variables for each stimulus) that are ignored by Worldlikeness during an experiment, but which are included in the downloadable results to aid in analysis.

textStimuli	audioStimuliPath	videoStimuliPath	imageStimuliPath	Session
Blick	Blick.wav	Blick.mp4	Blick.jpg	Practice
Bnick	Bnick.wav	Bnick.mp4	Bnick.jpg	Practice
Bwick	Bwick.wav	Bwick.mp4	Bwick.jpg	Formal
...	...	...	...	...

Table 3. A sample Worldlikeness stimulus list

Trial settings are customizable for each experiment in Worldlikeness, such as trial/stimulus duration and an inter-trial eye fixation cross ‘+’ with high precision at the millisecond level (depending on browser; see below), and response type (binary or seven-point Likert-scale response scales). During an experimental session, participants can respond by clicking a response button on the screen via mouse or touchscreen tap, or are required to respond by pressing a response key on the keyboard when the mouse cursor is set as hidden by experimenters in the experimental settings. All keys corresponding to letters (e.g., A-Z in English) and digits (i.e., 0-9) are available as response keys in Worldlikeness. Visual stimuli and the eye fixation cross can be set with a fixed size or a size proportional to the width of the screen, though they are always aligned to the center of the screen both vertically and horizontally.

Finally, to help prevent non-target Web users from participating in experiments, experimenters are offered a tool to create a four-way forced-choice language task as a linguistic competence pretest (Fig. 13). In a pretest, participants have three tries (10 seconds per try) to choose the correct answer. If they fail all three times, the system automatically blocks them from

participating in the experiment. When they pass the test, the time required to complete the task is also recorded as part of the downloadable participant information.

**Please answer the question below.**

Which of the following words is the homophone of 'bit'?

bitt  bight  bate  beet

Time left (sec): 8 Chances: 3

**Fig. 13** A four-way forced-choice English homophone test in Worldlikeness

*Data available to experimenters and researchers* – In Worldlikeness, experimenters and researchers can download four different types of experimental results files (Fig. 14). The ‘Participant’ file includes the participant IDs automatically assigned by Worldlikeness, the participant background information that experimenters choose to collect (e.g. age, gender, handedness, L1, L2, language impairment), experimental session starting/ending times, language test results, browser type, user interface (desktop vs. mobile) and IP address (this last available to corresponding experimenters only). The ‘Item’ file is the list of stimuli uploaded by experimenters plus the item IDs automatically assigned in Worldlikeness. The ‘Response’ file includes participant, item, and trial IDs automatically labelled in Worldlikeness, text stimuli, multimedia stimuli paths, durations of audio/video stimuli, participants’ responses, actual response keys, and reaction times. The ‘Exp Info’ files contains the experimental settings, consent form, instructions, and description. The first three files are in a comma-separated (CSV) format, and all files are compressed into a .zip file before being delivered to Worldlikeness users.



**Please select the data you would like to download.**

Participant[?]
  Item[?]
  Response[?]
  Exp Info[?]
  All

**Fig. 14** The dialogue box for downloading different types of experimental data

### **Reliability of data and reaction times in Worldlikeness**

To test whether Worldlikeness collects reliable response data and measure reaction times, we first ran a small-scale Mandarin wordlikeness judgment experiment in Worldlikeness (Authors, 2017) that replicated the megastudy reported in Myers (2015), which had been run in E-Prime (Schneider et al., 2002). In Myers’s megastudy, more than 3,000 nonwords were divided into two item sets and presented in Zhuyin Fuhao (the onset-rime-based phonetic orthography used in Taiwan) to more than 100 native speakers of Mandarin in Taiwan. The participants were asked to provide binary judgments on whether each of the 3,000 nonwords is like Mandarin or not. One major finding in Myers (2015) was that a nonce syllable was more likely to be accepted by native speaker of Mandarin if it had more phonological neighbors in Mandarin (differing in only one phoneme, ignoring tone).

To replicate this result in Worldlikeness, we randomly selected 100 items from each of the two nonword sets used in Myers’ megastudy, and created two Worldlikeness experiments accordingly. Eleven native speakers of Mandarin were recruited for the first experiment and twelve for the second one, who were asked to provide binary judgments (‘like Mandarin’ vs. ‘unlike Mandarin’) as in Myers (2015). The descriptive statistics for the two sets of judgment scores and reaction times (after removing trials without responses) are summarized in Tables 4 and 5,

respectively; the distributions of reaction times are visualized in Fig. 15. The probability of item acceptance was low for both item sets, but two-sample tests for equality of proportions still showed that acceptance in both was significantly higher than for the megastudy (Set 1:  $\chi^2(1) = 29.7, p < .001$ ; Set 2:  $\chi^2(1) = 150.6, p < .001$ ). Two-sample  $t$  tests of the log-transformed RTs suggest that participants also responded to both item sets significantly slower in our replication than in Myers (2015) (Set 1:  $t(1455) = 11.9, p < .001$ ; Set 2:  $t(1643) = 22, p < .001$ ).  $F$  tests also indicate a significant difference in variance between the two studies for both item sets (Set 1:  $F(1186) = 0.606, p < .001$ ; Set 2:  $F(1095) = 0.65, p < .001$ ).

Judgment – Set 1	Reject	Accept	Acceptance Prob.	Set 2	Reject	Accept	Acceptance Prob.
Authors (2017)	877	219	0.2		860	327	0.28
Myers (2015)	9,845	1,096	0.14		9,406	1,535	0.14

Table 4. Wordlikeness judgments in Authors (2017) and Myers (2015)

RT (ms) – Set 1	Min	Max	Mean (sd)	Set 2	Min	Max	Mean (sd)
Authors (2017)	197	3685	1055 (667)		13	3994	1231 (724)
Myers (2015)	1	3996	874 (642)		1	3999	872 (640)

Table 5. Reaction times (RT) in Authors (2017) and Myers (2015)



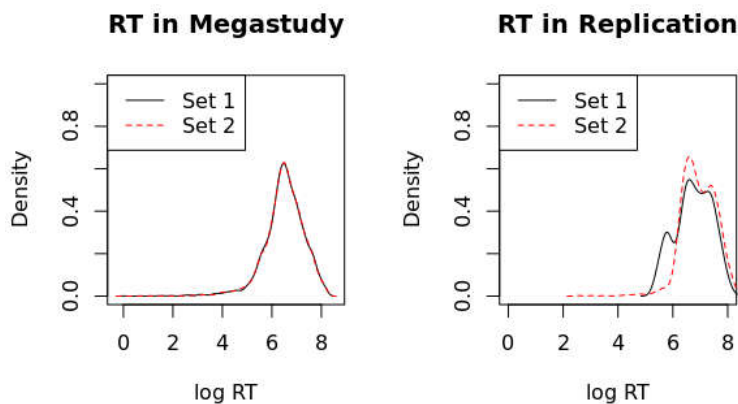


Figure 15. Distribution of reaction times (RT) of both item sets in Authors (2017) and Myers (2015)

The subjects were more likely to accept nonwords and responded more slowly in our replication due to a different experimental design in Myers (2015). Myers' megastudy included more than 3,000 items in the wordlikeness judgment task, in which “judgment fatigue” (Snyder, 2000) occurred after a long experimental session to gradually lower the acceptance rate and increase monotonous and thus faster responses. In sum, these discrepancies should not be ascribed to any fundamental issues in the design of Worldlikeness, and the same effect is expected to occur if our replication study using Worldlikeness has the same scale.

To further validate the similarity between the results in the two studies, by-item judgment score and reaction time means were compared across the two studies in a linear regression model. All reaction times shorter than 100 ms were likely due to erroneous key pressing before stimulus onset and were thus removed before the analyses (7 (0.3%) and 559 (2.6%) data points from our replication and Myers (2015), respectively). Reaction times were then log-transformed and averaged for each item within each item set in our replication and the Myers megastudy. In the linear regression models, mean judgment scores and reaction time  $z$ -scores from the megastudy

served as the sole predictor of their counterparts in our replication within each item set. As illustrated in Fig. 16, the mean judgment scores and reaction times from the megastudy was a significant predictor of those from both of our replications (Set 1 Judgment:  $B = 5.22$ ,  $SE = 0.73$ ,  $t = 7.15$ ,  $p < .001$ ; Set 1 RT:  $B = 0.59$ ,  $SE = 0.13$ ,  $t = 4.71$ ,  $p < .001$ ; Set 2 Judgment:  $B = 5.15$ ,  $SE = 0.73$ ,  $t = 7.07$ ,  $p < .001$ ; Set 2 RT:  $B = 0.48$ ,  $SE = 0.13$ ,  $t = 3.66$ ,  $p < .001$ ), though only relatively small proportions of variance were explained (Set 1 Judgment:  $r^2(98) = .343$ ,  $p < .001$ ; Set 1 RT:  $r^2(98) = .185$ ,  $p < .001$ ; Set 2 Judgment:  $r^2(98) = .338$ ,  $p < .001$ ; Set 2 RT:  $r^2(98) = .12$ ,  $p < .001$ ).

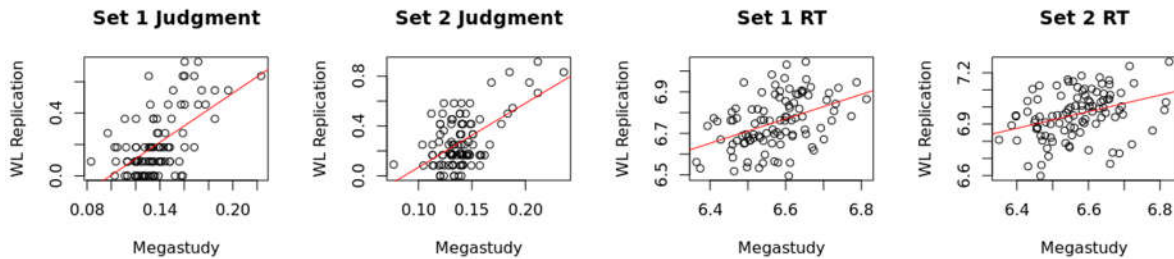


Figure 16. By-item comparisons of mean acceptability scores and log reaction times (RT) in Myers (2015) (megastudy using E-Prime) and Authors (2017) (replication using Worldlikeness)

We also analyzed the experimental results using maximal mixed-effect logistic regression (Bates et al. 2013) for both groups of data by including binary acceptability judgments as the dependent variable, log-transformed neighborhood density z-scores as the sole independent variable, and Participant ID and Item ID as random variables with their random slopes. The neighborhood density effect was significant for both item sets (Set 1:  $\beta = 0.77$ ,  $SE = 0.13$ ,  $z = 6.03$ ,  $p < .001$ ; Set 2:  $\beta = 0.56$ ,  $SE = 0.11$ ,  $z = 5.05$ ,  $p < .001$ ) (see the Appendix for links to these data sets). The strong positive correlations in wordlikeness judgment scores and reaction times between the E-Prime-run megastudy and our small-scale Worldlikeness-run experiments, as well as the

replication of the neighborhood density effect, provide an initial hint that judgment data collected via Worldlikeness may be as reliable as those from the widely used E-Prime software.

To evaluate the precision of reaction time measurement across Web browsers, we created a dummy Worldlikeness experiment with five trials in its practice session and eight trials in the formal session, and modified the client-side code in the Worldlikeness application to simulate a response key pressing event one second after the onset of every trial. The simulated key pressing client-side codes were built with the JavaScript function *setTimeout*, which delayed the execution of the codes that passed a key response event wrapped inside the function. In our tests, immediately before the key response function was initiated at the onset of a trial, the time on the participants' device was recorded as the onset time point. The execution of the key response function was set to delay for one second, and the time on the participants' device was immediately recorded again as the offset time point. The difference between the two time points was calculated as the reaction time of the trial. The dummy experiment was run ten times in each test browser, yielding 130 reaction time data points per browser. Differences in observed reaction times from the actual duration of 1000 ms were then analyzed as a function of browser type.

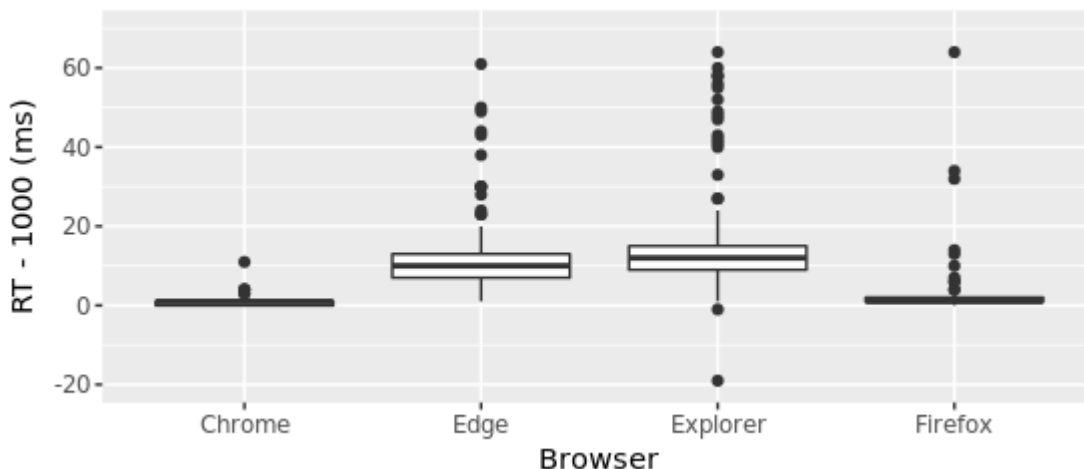
These benchmark tests were executed on a desktop computer with an AMD FX-8320 3.5 GHz octa-core CPU, eight gigabytes of DDR3 RAM, and Microsoft Windows 10 updated to ver. 1703. The four target major browsers included Microsoft Edge (ver. 40.15063), Microsoft Internet Explorer (ver. 11.296.15063), Mozilla Firefox (ver. 53.0.3), and Google Chrome (ver. 59.0.3071), which were updated to their latest version as of Jun 7, 2017. Microsoft Edge was chosen over Internet Explorer since it is Microsoft's latest highly promoted browser, and the last three browsers were chosen because they represent over 85% of the market share at the time of running the

benchmark tests (based on a survey retrieved from <https://www.netmarketshare.com/browser-market-share.aspx> on Oct 20, 2017).

The descriptive statistics of the benchmark tests is summarized in Table 6 and visualized in Fig. 17. The mean reaction time variation is notably higher in Microsoft Edge and Microsoft Internet Explorer than in Mozilla Firefox and Google Chrome, with the latter showing an impressive mean reaction time difference of less than one millisecond from the target of 1000 ms. A one-way ANOVA showed a significant effect of browser on reaction time variation values ( $F(3, 516) = 79.78, p < .001$ ). The results of post-hoc Tukey HSD pairwise comparisons can then be translated into a browser accuracy ranking: Google Chrome  $\approx$  Mozilla Firefox ( $p = .3$ ), Mozilla Firefox  $\gg$  Microsoft Edge ( $p < .001$ ), Microsoft Edge  $\gg$  Microsoft Internet Explorer ( $p < .01$ ).

Browser Type	Mean (sd)	Max	Min
Microsoft Edge	11.98 (9.61)	61	1
Microsoft Internet Explorer	15.65 (13.79)	64	-19
Mozilla Firefox	2.79 (6.92)	64	0
Google Chrome	0.84 (1.23)	11	0

Table 6. Descriptive statistics of reaction time variation values (ms) by browser



**Fig. 17** Differences in reaction time benchmarks across browsers; the top of a box = the 75<sup>th</sup> percentile, the bottom of a box = the 25<sup>th</sup> percentile, the band inside each box = median, upper whisker = the 75<sup>th</sup> percentile + 1.5\*interquartile range, lower whisker = the 25<sup>th</sup> percentile – 1.5\*interquartile range, black dots = outliers (Chrome: 4.6%, Edge: 10%, Explorer: 14.6%, Firefox: 8.5%)

There are two possible ways in which these cross-browser variations may arise: within WL specifically, or within JavaScript more generally. We are more inclined to attribute the variations to the performance of the client-side JavaScript codes within different Web browsers. Note that the key pressing simulation was implemented with the JavaScript function *setTimeout* in our benchmark tests. Different browsers, however, are known to vary in the precision of their time counter in the JavaScript function (e.g. Resig 2008), and thus differ in when a key pressing event is triggered after the set delay. This interpretation of the reaction time data is plausible since offset time points were recorded only *after* the delay in the *setTimeout* function was complete; the seeming imprecision in measuring reaction times is in fact due to the inaccurate automatic delay in *setTimeout* across Web browser. In any case, as noted earlier, Worldlikeness records participants'

browser information as part of the experimental data, so skeptical researchers can include it as a predictor in a statistical model to help factor out any potential effect of browser type on the reaction times recorded by Worldlikeness.

### **Wordlikeness judgments by Mandarin-Southern Min bilinguals**

As a test of the potential of Worldlikeness, we seeded it with wordlikeness data collected for a study of lexical and social influences on the acceptability of monosyllabic nonwords in bilingual speakers of Mandarin and Taiwan Southern Min (commonly called Taiwanese). These two major languages spoken in Taiwan are interesting for a small-scale cross-linguistic study for a number of reasons. First, while both are members of the Sinitic language family, they still differ considerably and are thus not mutually intelligible. For example, there are fewer coda consonants, lexical tones, and lexical monosyllables in Mandarin than in Southern Min. Moreover, the logographic writing system has a long history in the development of modern Mandarin, whereas the education system in Taiwan did not introduce any official writing system for Southern Min until the 1990s, and it is still not widely known, let alone used, by Southern Min speakers. A second reason for looking at these two languages is that many speakers in Taiwan are Mandarin/Southern-Min bilinguals, raising additional cross-lexical/speaker-internal issues (see, e.g., Lemhöfer et al., 2008). Third, the two languages also differ in social status in Taiwan, with Mandarin being more prestigious than Southern Min. We thus expect to see effects of social variables on language processing consistent with previous research (e.g., female speakers may tend to favor the prestige norm; Labov, 2001), and possibly interactions between the social and lexical/cognitive/phonological variables as well. Accordingly, the current experiments were

designed as auditory wordlikeness judgment tasks, not only because Southern Min is generally not written, but also because the variable of speaker accent could only be examined in the auditory modality.

While we are still exploring these theoretical issues experimentally, the methodological aspects of this study in progress also highlight important features of Worldlikeness. In particular, we show here that (i) Worldlikeness allows experiments to be run in different modalities, (ii) Worldlikeness makes it efficient to recruit participants for disparate target languages online, and (iii) differences in the results from in-lab vs. online participation are minor and can be factored out.

### *Method*

*Materials* – We used all onset and rhyme types, plus lexical tones, in Mandarin and Southern Min to generate all possible combinations following the basic syllable template (C)(G)V(G/C), shared by both languages (C = consonant, V = vowel, G = glide). Lexical syllables in Mandarin and Southern Min were then excluded from this list of monosyllables. From these more than 5,000 non-lexical syllables, we further excluded those created with a mid-level tone, phonemic only in Southern Min, to avoid perceptual confusions when judging syllables in Mandarin. Syllables with an obstruent coda are also possible only in Southern Min, so they were also left out as too clear a violation of Mandarin phonotactics.

Among the remaining nonlexical syllables, we randomly selected 200 items to use. We presented the stimuli in IPA to two female lab assistants whose home languages were, respectively, Mandarin (speaker KY) and Southern Min (speaker PS), and asked them to read each stimulus aloud. There was no specific instruction unless the speakers had any difficulties producing the syllables naturally, in which case the first author demonstrated the pronunciation of the syllables.

Recordings were made in a sound-attenuated room using Praat (Boersma & Wernicke, 2017) with a sampling rate at 44,100 Hz.

A screening process was applied to exclude the 71 items produced by both speakers that were judged by twelve Mandarin-Southern Min bilinguals as real either in Mandarin or Southern Min. This process was expected to further lower the chance of perceptual confusion between nonlexical monosyllables and real words during the auditory wordlikeness judgments.

For our calculations of the neighborhood density (ND) for each test item in each target language, we first extracted all unique Mandarin monosyllables presented in Zhuyin Fuhao from Tsai's (2000) full list of Mandarin characters, and then converted Zhuyin Fuhao spellings into IPA transcriptions and corresponding ASCII codes in which each phoneme was represented by one unique symbol. The Mandarin ND of each item was calculated as the number of lexical neighbors different in exactly one segment, ignoring tones. We then translated all IPA transcriptions of all lexical monosyllables in Southern Min from Dong (2001) into ASCII codes using the same system so each ASCII code in the two languages represented the same phoneme. The Southern Min NDs were calculated in the same way.

*Participants* – 81 bilingual speakers of Mandarin and Southern Min (55 males and 26 females) enrolled as undergraduates or graduate students in southern Taiwan were recruited to participate in the wordlikeness judgment tasks. The age of the in-lab participants ranged from 18 to 28 years (mean = 21.6, sd = 1.95). Another 156 bilingual speakers (103 males, 52 females, 1 transgender) were recruited by a Web advertisement posted on the internet via Facebook in less than two weeks,



and participated in the experiments online.<sup>1</sup> The self-reported ages of the online participants ranged from 18 to 62 years (mean = 26.8, sd = 7.6), which was significantly higher ( $t(191) = 8.1$ ,  $p < .001$ ) and more variable ( $F(155) = 15.14$ ,  $p < .001$ ) than those of the in-lab participants. In-lab participation was compensated with NT\$50, and both groups of participants were rewarded with a result report within Worldlikeness after the end of their experimental session (see the ‘The Worldlikeness application design’ section above).

*Procedure* – The in-lab and online participants were randomly assigned to one of the four conditions (2 Accent  $\times$  2 Target Languages) in Worldlikeness. Both groups of participants had to pass a Southern Min language proficiency task in Worldlikeness (see the ‘The Worldlikeness application design’ section above) in order to proceed to the experimental session. In the four-way forced-choice task, participants were instructed to choose a lexical word in Southern Min containing a specified ‘sound’ (phoneme). The in-lab participants were asked to complete an additional task by reading aloud a written Mandarin paragraph in Southern Min to confirm their fluency in the target language. At the onset of each trial in the experimental session, a random auditory stimulus was selected without replication and presented to the participants, who were asked to judge whether the nonword monosyllables sounded like the target language or not. The participants were given four seconds from the onset of the trial to respond by pressing ‘S’ or ‘L’ on the keyboard for ‘unlike’ and ‘like’ respectively. After a response was provided, or no response was given within the four-second time limit, the experimental session proceeded automatically to the next trial. A one-second frame with a horizontally and vertically aligned eye fixation cross ‘+’ was inserted between every two trials.

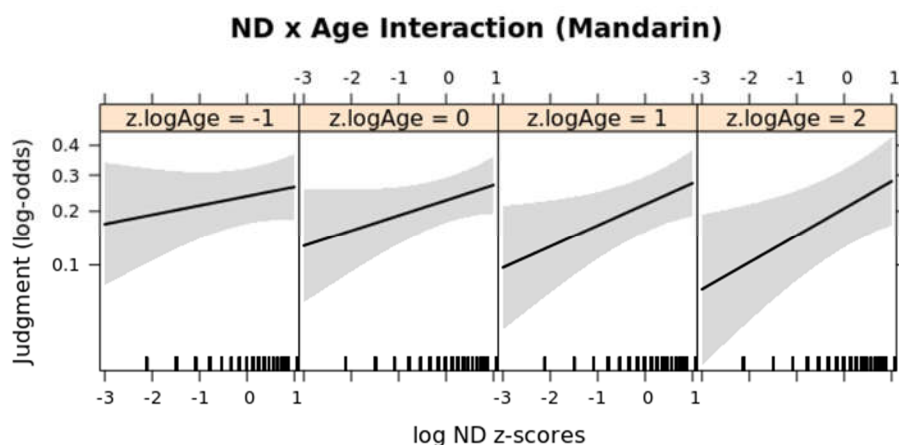
---

<sup>1</sup> The Facebook post <https://www.facebook.com/lingproc.exp/posts/1355912174498901> has reached 13,122 people as of Oct 1, 2017.

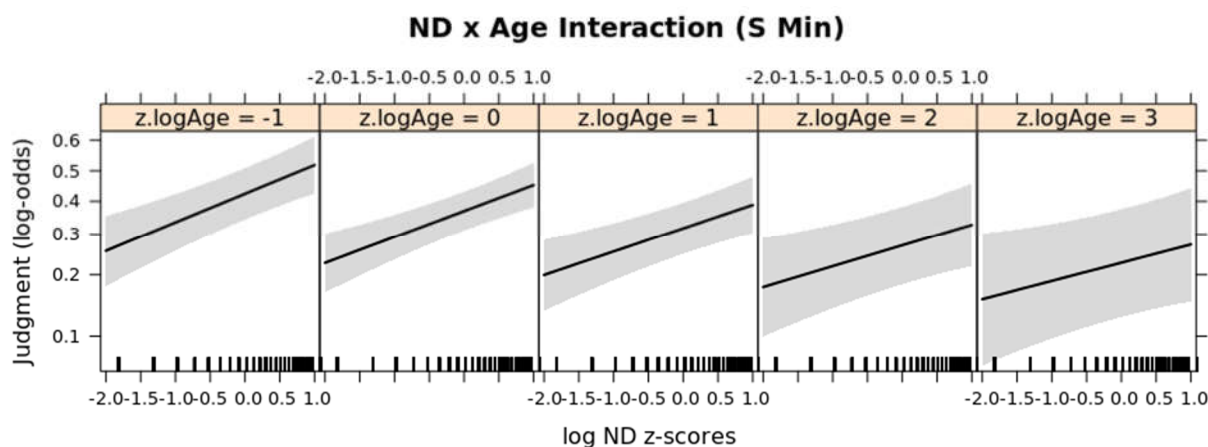
*Results* – To focus the discussion here on the methodological issues, we selected results from participants in the conditions where the target language and accent were consistent. Accordingly, the Mandarin data set included 19 in-lab participants (age = 18-26, mean = 21.8, sd = 2.1) and 39 online participants (age = 18-43, mean = 26.9, sd = 6), and the Southern Min data set had 21 in-lab participants (age = 19-26, mean = 21, sd = 1.6) and 40 online participants (age = 18-54, mean = 26.7, sd = 7.9). ND and age were log-transformed and z-scored within the two data sets before starting the statistical analysis. The two sets of experimental results were analyzed separately using mixed-effect logistic regression (Bates et al., 2012), with wordlikeness judgment as the dependent variable. Mandarin ND or Southern Min ND (depending on the target language), Setting (i.e., in-lab vs. online participation), and Age were the three independent variables, along with the two-way ND  $\times$  Setting and ND  $\times$  Age interactions. Age was included because of the substantial difference in age between the in-lab and online groups, and we wanted to disentangle effects of Setting (relevant to the validity of online experimentation) from Age (an orthogonal factor). Test item ID and participant ID were included as random variables (intercepts only, to allow for model convergence; Matuschek et al., 2017).

The effect of Mandarin ND on Mandarin wordlikeness judgment was only marginally significant ( $\beta = 0.27$ ,  $SE = 0.15$ ,  $z = 1.82$ ,  $p = .07$ ), but this trend is consistent with previous findings in Myers (2015) and Authors (2017). The ND  $\times$  Age interaction was significant ( $\beta = 0.09$ ,  $SE = 0.05$ ,  $z = 1.94$ ,  $p = .05$ ), suggesting that older speakers were more influenced by the number of phonological neighbors in their lexicon (Fig. 18). There was no significant effect of Setting ( $z = -1$ ;  $p = .32$ ) nor significant ND  $\times$  Setting interaction ( $z = -0.58$ ,  $p = .56$ ), which indicates similar judgment patterns for both in-lab and online participants, after Age was taken into account. The

Southern Min data also showed a significant positive ND effect ( $\beta = 0.42$ ,  $SE = 0.09$ ,  $z = 4.9$ ,  $p < .001$ ) as well as a significant negative Age effect ( $\beta = -0.23$ ,  $SE = 0.11$ ,  $z = -2.03$ ,  $p < .05$ ) (Fig. 19). The latter suggests that older participants were less likely to accept nonwords in Southern Min wordlikeness judgments. Again, there was no significant effect of Setting, but this time also no significant interactions ( $z_s < -1.75$ ;  $p_s > .08$ ).



**Fig. 18** Mandarin ND  $\times$  Age interaction



**Fig. 19** Southern Min ND  $\times$  Age interaction

*Discussion* – The results were consistent with previous findings that neighborhood density plays an important role in the online visual processing of phonological forms. In addition, our study validates the online administration of wordlikeness experiments in Worldlikeness since there was no significant effect of Setting per se. Instead, we found an effect of Age, as the online experiments reached a population with an overall higher age and greater age variation than the in-lab study on college studies. The stronger positive effect of neighborhood density on Mandarin worldlikeness judgments for older participants might be attributed to their larger vocabulary size (e.g., Keuleer et al., 2015; Meylan & Gahl, 2014) and thus different neighborhood density measures; more acceptable nonwords might have had a greater number of mentally accessible phonological neighbors for older speakers. The negative correlation between nonword acceptability and age in Southern Min wordlikeness judgment could have been due to the fact that the accent of our native Southern Min speaker (PS), being younger than many of the online participants and under the influence of a major language shift (e.g., Young, 1998). A sociolinguistic survey conducted by Chen (2004) showed that the fluency of Taiwanese local languages, including Southern Min, has dropped significantly for speakers of younger generations. This change may have contributed to lower the Southern Min acceptability of the speaker’s nonwords, as judged by older speakers, whose dominant language is likely to be Southern Min (25 out of 39 online participants of 30 years old or older in our study reported Southern Min as their mother tongue). Regardless of how these variables sort themselves out in further studies, we have shown that despite the partial confounding between age and experimental setting, the factors can be teased apart. Moreover, for studies explicitly focused on speakers beyond the usual college-aged participants, it is good to know that such speakers are readily accessible online.

## General discussion

In this paper, we have highlighted the importance of studying typological psycholinguistics via web crowdsourcing, and showed how Worldlikeness was developed to help expand the research scope, reduce the overt collaborative burden, and recruit native speakers of different languages online. Worldlikeness judgment tasks were designed and administered in different modalities (visual vs. auditory) in separate settings (in-lab vs. online) via Worldlikeness not only to replicate previous findings on monolingual phonological processing, but also to provide new insights into cross-linguistic and multilingual phonological processing. We are currently taking advantage of the convenience and reliability of Worldlikeness to extend our cross-linguistic experiments and analyses to include other Sinitic languages such as Hakka as well as Taiwan Sign Language and seek to further explore the universal and idiosyncratic nature of phonological processing. We are also working to improve Worldlikeness still further to accommodate the complex needs of typological psycholinguistics. Here we discuss two of the improvements currently in progress.

*Automatized generation of nonwords and calculation of lexical variables* – We are working to equip Worldlikeness with a tool to increase methodological consistency across experiments created and run in Worldlikeness. This is a function that can generate nonword test items using the same algorithm for each target language from an electronic dictionary uploaded by experimenters, similar to what WordGen (Duyck et al., 2004) does for the small set of languages that its algorithms were designed for. Our tool will help avoid possible biases in the selection of nonwords across languages. Relatedly, we hope to allow Worldlikeness to calculate lexical variables for automatically generated nonce words to avoid variation in their quantification across research

groups. For example, neighborhood density is traditionally defined as the number of lexical words that differ from the target word in exactly one segment (e.g., Luce & Large, 2001), but for longer words it is sometimes quantified as mean phonological Levenshtein (edit) distance from the twenty nearest lexical neighbors (i.e., PLD20) in Yarkoni et al. (2008). While it is possible to standardize different variable scales (e.g., via *z*-score transformation), with implementation of these automatized procedures in Worldlikeness, meta-analyses of cross-linguistic wordlikeness results would become more straightforward.

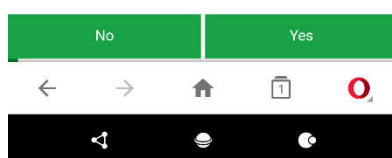
*Browser compatibility on mobile devices* – For a comprehensive typological psycholinguistic survey, it is crucial to study languages that vary substantially in their linguistic characteristics. Some typologically important but understudied languages are spoken or signed in developing countries where mobile networks are better established than fixed networks via telephone or cable lines (e.g., Aker & Mbiti, 2010). To make Worldlikeness more accessible to people in these areas, we have developed a user interface optimized for mobile device. The main difference between the desktop and mobile user interfaces is the touch-friendly layout in the latter, with which experimenters can quickly manage their experiments and participants can easily respond in an experimental session via their mobile devices.

Currently, the mobile interface of Worldlikeness is ready for running experiments using text stimuli in major Android browsers (e.g., Google Chrome, Mozilla Firefox, Opera; Fig. 20). However, since mobile browsers can behave very differently in dealing with Web multimedia files as compared to their desktop counterparts, we are still testing different user interfaces to provide a uniform presentation of multimedia stimuli across different mobile browsers. The mobile interface shares the core JavaScript functions with the desktop interface, and records judgment

data and measures reaction times in exactly the same way. Therefore, we expect judgment data collection to be as accurate in Worldlikeness and reaction times to be influenced by similar environmental variables. That said, we will still benchmark the reaction time measures and test different layouts for the consent form page in the mobile user interface to motivate participants to share their experimental data publicly in order to ultimately confirm the cross-browser compatibility of Worldlikeness .



civer



**Fig. 20** A sample English wordlikeness judgment with text stimuli in the mobile user interface in the Android Opera browser

In sum, it is our hope that the emphasis on the Web crowdsourcing features in Worldlikeness will attract more linguists and psychologists to run experiments on typologically

distinct languages and share their data online, or even inspire the creation of new Web-based tools incorporating similar concepts as we have demonstrated here.

## References

- Aker, J. C., & Mbiti, I. M. (2010). Mobile phones and economic development in Africa. *Journal of Economic Perspectives*, 24(3), 207–232.
- Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1), 9–41.
- Authors (2017).
- Baayen, R. H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, 5(1), 149–157.
- Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44(4), 568–591.
- Balota, D. A., Yap, M. J., Hutchison, K. T., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual Word Recognition, Vol. 1: Models and methods, orthography, and phonology* (pp. 90–115). London, UK: Psychology Press.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, 133(2), 283–316.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing : Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.



- Bates, D., Bolker, B., Maechler, M., & Walker, S. (2013). *lme4: Linear mixed-effect models using Eigen and Eigen++*. R package version 0.999999-2. Retrieved from <http://cran.r-project.org/web/packages/lme4/index.html>
- Bates, E. (1999). Processing Complex Sentences: A Cross-linguistic Study. *Language and Cognitive Processes*, 14(1), 69–123.
- Bates, E., D’Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., ... Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, 10(2), 344–380.
- Boersma, P., & Weenick, D. (2017). *Praat: doing phonetics by computer* (Version 6.0.35) [Computer program]. <http://www.fon.hum.uva.nl/praat/>. Retrieved Oct 20, 2017.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1), 3–5.
- Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Word-Form Encoding in Mandarin Chinese as Assessed by the Implicit Priming Task. *Journal of Memory and Language*, 46(4), 751–781.
- Chen, S. C. (2004). *Linguistic Vitality in Taiwan: A Sociolinguistic Study*. Research project report, Ministry of Science and Technology, Taiwan.
- Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *Computational phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology* (pp. 49–56). Somerset, UK: Association for Computational Linguistics.
- Coleman, T., & Greif, S. (2016). *Discover Meteor*. <https://discovermeteor.com> (retrieved on Jun 19, 2017).

- 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.
- Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10, 65-70.
- de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1), 1–12.
- Dong, Z. S. (2001). *Tai Wan Min Nan Yu Tsi Dian [The Dictionary of Taiwanese Southern Min]*. Taipei: Wu-nan Culture Enterprise.
- Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, 36(3), 488–499.
- Erlewine, M. Y., & Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory*, 34(2), 481–495.
- Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S. and Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, 22, 421-435.
- Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of Wordlikeness: Effects of Segment Probability and Length on the Processing of Nonwords. *Journal of Memory and Language*, 42(4), 481–496.
- Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, 66(3), 363–376.

- 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.
- Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, 59(2), 93–104.
- Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2), 366–385.
- Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America*, 107, 2711–2724.
- Han, M. S. (1994). Acoustic manifestations of mora timing in Japanese. *The Journal of the Acoustical Society of America*, 96(1), 73–82.
- Harrell, F. E. Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis, 2<sup>nd</sup> edition*. New York: Springer-Verlag.
- Haspelmath, M., Dryer, M.S., Gil, D., & Comrie, B. (Eds.) (2005). *The world atlas of language structure*. Oxford: Oxford University Press.
- Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, 44(1), 45–75.
- Hyman, L. M. (1985). *A Theory of Phonological Weight*. Dordrecht: Foris. Jaeger, T. F., & Norcliffe, E. J. (2009). The Cross-linguistic Study of Sentence Production. *Language and Linguistics Compass*, 3(4), 866–887.

- Jarema, G., Busson, C., Nikolova, R., Tsapkini, K., & Libben, G. (1999). Processing Compounds: A Cross-Linguistic Study. *Brain and Language*, 68(1–2), 362–369.
- Johnson, E. J., Ballman, S., & Lohse, G. L. (2002). Defaults, framing, and privacy: Why opting in  $\neq$  opting out. *Marketing Letters*, 13(1), 5–15.
- Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods*, 41(1), 1–12.
- Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology*, 68(8), 1457–1468.
- Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, 68(8), 1665–1692.
- Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. *The Quarterly Journal of Experimental Psychology*, 68(8), 1693–1710.
- Kirby, J. P., & Yu, A. C. L. (2007). Lexical and phonotactics effects on wordlikeness judgments in Cantonese. In *Proceedings of the International Congress of the Phonetic Sciences XVI* (pp. 1389–1392).
- Labov, W. (2001). *Principles of Linguistic Change, Vol. 2: Social Factors*. Oxford, UK: Blackwell.
- Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 12–31.

- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2), 433–442.
- Löfgren, A., Martinsson, P., Hennlock, M., & Sterner, T.(2012). Are experienced people affected by a pre-set default option – Results from a field experiment. *Journal of Environmental Economics and Management*, 63, 66–72.
- Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, 16(5–6), 565–581.
- Mantonakis, A., Rodero, P., Lesschaeve, I., & Hastie, R. (2009). Order in choice: Effects of serial position on preferences. *Psychological Science*, 20(11), 1309–1312.
- McCarthy, J., & Prince, A.. (1986). Prosodic morphology. Ms., University of Massachusetts, Amherst.
- Meylan, S., & Gahl, S. (2014). The Divergent Lexicon: Lexical Overlap Decreases With Age in a Large Corpus of Conversational Speech. *Proceedings of the Cognitive Science Society*, 36(36).
- Miller, J. M., & Krosnick, J. A. (1998). The impact of candidate name order on election outcomes. *The Public Opinion Quarterly*, 62(3), 291–330.
- MongoDB, Inc (2008-2017). mongoDB (Version 3.4) [Computer program]. <https://mongodb.com>. Retrieved on Oct 20, 2017.
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen H. R., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315.
- Myers, J. (2016). Meta-megastudies. *The Mental Lexicon*, 11(3), 329–349.

- Myers, J. (2015). Markedness and Lexical Typicality in Mandarin Acceptability Judgments. *Language and Linguistics*, 16(6), 791–818.
- Myers, J. (2016). Meta-megastudies. *The Mental Lexicon*, 11(3), 329–349.
- Norcliffe, E., Harris, A. C., & Jaeger, T. F. (2015). Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances. *Language, Cognition and Neuroscience*, 30(9), 1009–1032.
- O’Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, 115(2), 282–302.
- Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or Syllable? Speech Segmentation in Japanese. *Journal of Memory and Language*, 32(2), 258–278.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), 411–419.
- Park, C. W., Jun, S. Y., & MacInnis, D. J. (2000). Choosing what I want versus rejecting what I do not want: An application of decision framing to product option choice decisions. *Journal of Marketing Research*, 37(May), 187–202.
- Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2, 10.
- Resig, J. (2007). Analyzing timer performance. <https://johnresig.com/blog/analyzing-timer-performance/> (retrieved on Jun 13, 2017)
- Ryan, K. J., Brady, J. V., Cooke, R. E., Height, D. I., Jonsen, A. R., King, P., Lebacqz, K., Louisell, D. W., Seldin, D. W., Stellar, E., & Turtle, R. H. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Washington,

- DC: National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime: User's guide*. Pittsburgh, PA: Psychology Software Incorporated.
- Snyder, W. (2000). An Experimental Investigation of Syntactic Satiation Effects. *Linguistic Inquiry*, 31(3), 575–582.
- Strange, W. (Ed.). (1995). *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York Press.
- von Bastian, C. C., Locher, A., & Rufin, M. (2013). Tootool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods*, 45(1), 108–115.
- Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, 58(2), 250–271.
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5), 971–979.
- Zhou, X., & Marslen-Wilson, W. D. (2000). Lexical representation of compound words: Cross-linguistic evidence. *Psychologia*, 43(1), 47–66.

### **Appendix. Links to experimental data sets**

Megastudy data sets in Myers (2015)

<http://lngproc.ccu.edu.tw/MWP/syllableJudgements.html>

Experimental results of the first stimulus group in the replication study of Myers (2015)

<https://www.worldlikeness.org/#/resultsInfo/EbP6EmD9vPiYHYXSd>

Experimental results of the second stimulus group in the replication study of Myers (2015)

<https://www.worldlikeness.org/#/resultsInfo/poMyogBwDWY2roRZC>

Results of wordlikeness judgment tasks

Condition 1: Target Language = Mandarin, Speaker = KY (Mandarin Accent), Authorization

Option Layout = Open First

In-lab participation: <https://www.worldlikeness.org/#/resultsInfo/TvmjnxSuNpTpBuG5M>

Online participation: <https://www.worldlikeness.org/#/resultsInfo/hvr7cRFsLeEEb9Yh4>

Condition 2: Target Language = Mandarin, Speaker = PS (Southern Min Accent), Authorization

Option Layout = Open First

In-lab participation: <https://www.worldlikeness.org/#/resultsInfo/aso5LqnkNpK5Et3hC>

Online participation: <https://www.worldlikeness.org/#/resultsInfo/DBtjdguzXgeXSZNAi>

Condition 3: Target Language = Southern Min, Speaker = KY, Authorization Option Layout =

Open First

In-lab participation: <https://www.worldlikeness.org/#/resultsInfo/et8eSa5c4wLv2dQbY>

Online participation: <https://www.worldlikeness.org/#/resultsInfo/LPgnLPXZ2YiQBjxxs>



Condition 4: Target Language = Southern Min, Speaker = PS, Authorization Option Layout = Open First

In-lab participation: <https://www.worldlikeness.org/#/resultsInfo/No2ATWirJPoeN5ZDZ>

Online participation: <https://www.worldlikeness.org/#/resultsInfo/JA8n3i2QvCCb439g6>

Condition 5: Target Language = Mandarin, Speaker = KY, Authorization Option Layout = Open Last

In-lab participation: <https://www.worldlikeness.org/#/resultsInfo/yACHq5Et9fksc6Mfs>

Online participation: <https://www.worldlikeness.org/#/resultsInfo/mcbDED4JPpXLqDLyC>

Condition 6: Target Language = Mandarin, Speaker = PS, Authorization Option Layout = Open Last

In-lab participation: <https://www.worldlikeness.org/#/resultsInfo/M22ihFbk3R9AhA6Ma>

Online participation: <https://www.worldlikeness.org/#/resultsInfo/TqqvJGeZ6wi99x3hq>

Condition 7: Target Language = Southern Min, Speaker = KY, Authorization Option Layout = Open Last

In-lab participation: <https://www.worldlikeness.org/#/resultsInfo/ncdxSvpkisrqS6Go6>

Online participation: <https://www.worldlikeness.org/#/resultsInfo/WwHvxqdr7Y6W7voWR>

Condition 8: Target Language = Southern Min, Speaker = PS, Authorization Option Layout = Open Last

In-lab participation: <https://www.worldlikeness.org/#/resultsInfo/Qdr8LpmB6ym4DiLAp>

Online participation: <https://www.worldlikeness.org/#/resultsInfo/3BYBHaPttKu3EeAzN>