Worldlikeness: A Web crowdsourcing platform for typological psycholinguistics

Tsung-Ying Chen
Department of Foreign Languages and Literature, National Tsing Hua University, Hsinchu, Taiwan

James Myers
Graduate Institute of Linguistics, National Chung Cheng University, Chiayi, Taiwan

Abstract Worldlikeness, hosted online at https://worldlikeness.org, is a free Web crowdsourcing experimental tool and database that seeks to make it easier to generate and share cross-linguistic processing data. Adopting HTML5 technology and a responsive Web design using the JavaScript-based Meteor package, Worldlikeness supports the use of text and multimedia stimuli in a user interface optimized for participants using either desktops or mobile devices. Worldlikeness allows individual psycholinguists to collect data on individual languages, store the data in the Worldlkeness server, and share them to help build a cross-linguistic database. As the database grows, outside researchers can freely 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 introduce the key features of Worldlikeness and compare online wordlikeness judgments to in-lab data to demonstrate its reliability.

Keywords typology, psycholinguistics, wordlikeness, Web crowdsourcing, responsive Web design

1 Introduction

Online crowdsourcing has become an ever more popular data source in linguistics (e.g., Enochson and Culbertson 2015; Sprouse 2011; Keuleers et al. 2015). The large amount of data that can be collected online has increased the feasibility of the megastudy approach (Balota et al. 2012), in which researchers model language processing in terms of many partly confounded variables that do not lend themselves to traditional factorial designs. A growing collection of Web-based tools, some specifically designed for linguistic research, has facilitated the running of online experiments, including Amazon Mechanical Turk (Paolacci et al. 2010), tatool (von Bastian et al. 2013), WebExp (Keller et al. 2009), and Pavlovia (https://pavlovia.org/; Peirce and MacAskill 2018). Some of these tools also take advantage of the rapid development of mobile technology. Running online linguistic experiments on mobile devices has major benefits: more people respond to online surveys with their mobile devices than with their desktop PCs (e.g., Antoun et al. 2017), and in some developing countries, mobile networks are better established than fixed networks via telephone or cable lines (e.g., Aker and Mbiti 2010).

However, one application of online experimentation remains underexplored: typological psycholinguistics (Norcliffe et al. 2015). Structurally different languages are processed differently. For example, while all languages presumably have separate levels for phonemes and syllables (e.g., McCarthy and Prince 1996), native speakers of different languages do not weight these levels equally: Mandarin Chinese speakers treat the syllable as most important (e.g., Chen et al. 2002, O'Seaghdha et al. 2010), while English speakers are more sensitive to the phoneme level (e.g., Frisch et al. 2000, Hayes and White 2013, O'Seaghdha et al. 2010). Since any two languages differ in many ways (e.g., Chinese and English differ not only in orthographic systems but also syllable inventory sizes), a much larger sample of languages is needed for statistical methods to determine how cross-linguistic factors actually affect processing, in what Myers (2016) dubs meta-megastudies.

Online crowdsourcing makes meta-megastudies more feasible by providing access to a wider variety of languages than could be tested in any single lab. However, typological research also depends on data sharing and methodological consistency (e.g., https://wals.info/, Dryer and Haspelmath 2013). While there has been a welcome move towards open data repositories in psycholinguistics (e.g., Pavlovia, mentioned above, and the ManyBabies project at https://osf.io/rpw6d/; Frank et al. 2017), studies on different languages may use entirely different methods, complicating typological analyses.

This is the challenge that we intend to address with our free online experimental tool *Worldlikeness*. In addition to the mobile-user-friendly interface that increases the number of participants and languages that can be tested, Worldlikeness is designed to encourage independent psycholinguists to share their results with typological goals in mind, all while giving participant privacy and autonomy the utmost priority. This design reflects our belief that the key to typological psycholinguistics lies not in ever-larger research teams but in the sharing of independently collected, but methodologically constrained, data sets on individual languages.

2 Worldlikeness in a nutshell

Worldlikeness is a ready-to-use Web application developed using the JavaScript-based programming language Meteor (https://www.meteor.com) integrated with the server-side database package MongoDB (https://www.mongodb.com/). In this framework, Worldlikeness always follows the most updated and uniform Web programming standards (HTML5 and ECMA script) and is thus fully compatible with major modern desktop and mobile Web browsers (e.g., Apple

Safari, Google Chrome, Mozilla Firefox, and Microsoft Internet Explorer and Edge; Figure 1). The experimenter and researcher interfaces are currently available in English and Chinese, and the participant interface can be changed by uploading a translation script in any language. Worldlikeness fits both desktop and mobile use, adjusting its user interface depending on the width of the screen resolution (i.e., responsive Web design): a width under 1,000 pixels is likely a mobile device in its portrait view and thus changes the user interface into a mobile-friendly setting (Figure 1; right). We consider this purely Web-based approach a more preferable choice than programming different versions of smartphone apps for different mobile operating systems, with great advantages for the future development of Worldlikeness and its usability. The simple Web design allows us and future developers to update just the Web pages to guarantee the same functionality in *both* desktop and mobile settings. We hope this unified Web design will help Worldlikeness gain in popularity as an open-source application in the future, as it is easier for other researchers to follow the standard Web design and modify Worldlikeness for their own use. In addition, users of Worldlikeness need no additional apps, plug-ins, programming, or server management skills to create, manage, run, share, and participate in Web experiments. Worldlikeness is also a lightweight Web application, as it requires users to download only a total of 833.8 KB of compressed data under an SSL-encrypted connection,[1] and does not impose the extra burden on users to install yet another sizable phone app. Worldlikeness thus serves as a true cross-platform application optimized for mobile Web browsing.

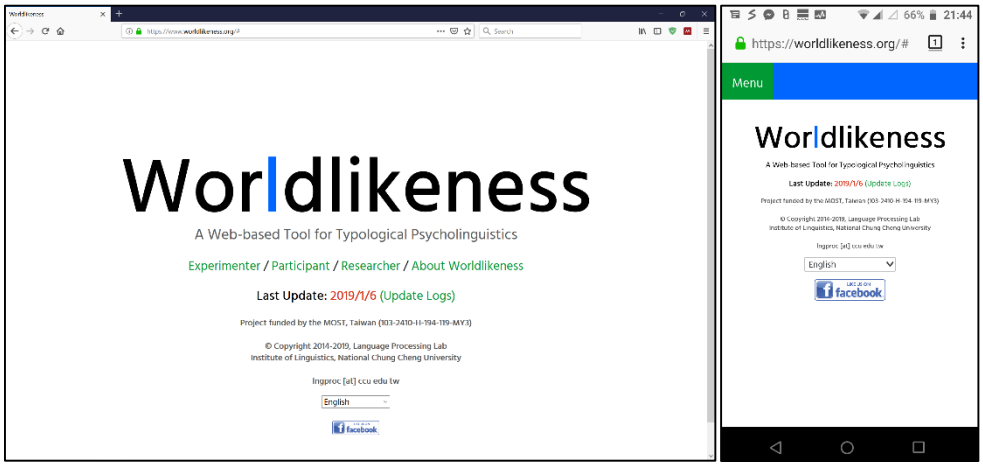

Figure 1. Worldlikeness front page in Mozilla Firefox (left: v64 in Windows 10 64-bit; right: v64 in Android 8.0)

Worldlikeness was originally designed for studying wordlikeness (the intuitive acceptability of non-words), hence the punning name, but it can accommodate any experiment with a simple stimulus-response trial structure (e.g., non-primed lexical decision tasks, perceptual discrimination tasks, tasks requiring the detection of images, words, syllables, or phonemes, and other subjective judgment tasks like semantic relatedness, concreteness, and so on; see Section 4 for planned future updates that incorporate some common experimental paradigms). Experimenters can present text, audio, image, and video stimuli and collect response choices as

[1] The application includes 686 KB of JavaScript files and 122 KB of an image file, which load in 1.9 s; test run on Jan 7, 2019 at https://tools.pingdom.com/.

button presses, mouse clicks, or screen taps, and record reaction times using the internal clock of the participant's device. Figure 2 illustrates the presentation in Android Mozilla Firefox of text, image, and video stimuli as could be used in testing written English non-words, nonlexical Chinese characters (see Myers 2019), and natural sign languages of the Deaf (see Lee 2016; image from Tsay et al. 2017). All stimuli are always downloaded to the users' device prior to an experimental session to ensure stability during data collection. A demo English wordlikeness judgment experiment can be accessed in both desktop (Front Page → Experimenter → Demo) and mobile Web browsers (Menu → Demo).

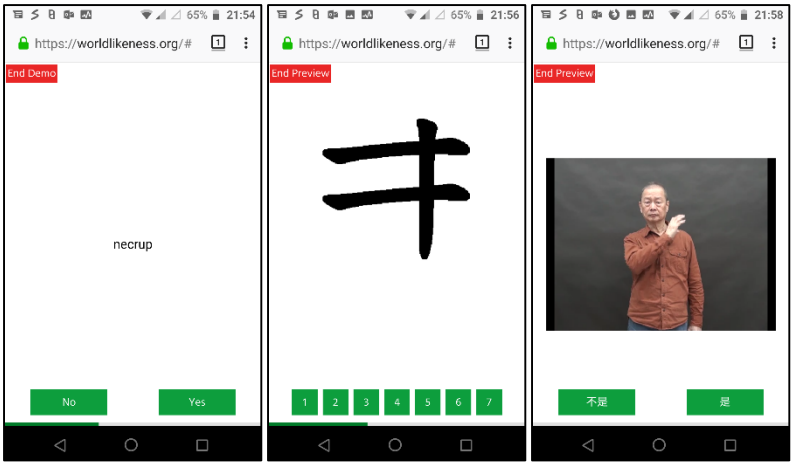

Figure 2. Presenting text (left), image (middle), and video (right) stimuli in mobile Mozilla Firefox v64 in Android 8.0; participants respond by tapping a green response button near the bottom

There are three user roles in Worldlikeness. *Experimenters* create, run, and share online experiments, *participants* are speakers/signers of a target language recruited by experimenters, and *researchers* download and analyze data shared by the experimenters. We have put great effort into developing a system that conforms to the core ethical principles in the Belmont report (Ryan et al. 1979). Thus only experimenters need to register, providing only minimal personal information (i.e., an e-mail to confirm identity) and a Web link in their experiments for participants to check the bona fides of experimenters and contact them with questions or complaints. All users, including participants, have complete control over their own data.

In the desktop interface of Worldlikeness, experimenters are provided with common customizable experimental settings for simple linguistic judgment tasks, particularly those crucial for wordlikeness studies: trial duration, response keys (A-Z and 0-9), visual stimulus size, an eye-fixation cross, and stimulus presentation frequency (if sampling from a larger stimulus set). Experimenters can also choose to collect basic participant information (e.g., age and gender) or require a brief language proficiency pretest. Worldlikeness records browser type and screen resolution for experimenters to take into account when comparing results from desktops and mobile devices. Experimenters wanting to test a large number of items also have the option of presenting a small randomly selected subset to each participant. Latin square designs are also possible by grouping experiments together. While experimenters are welcome to keep their experimental results private, sharing experiments increases their quota for further experiments. Worldlikeness also provides direct links to publicly shared results, which makes it easier for experimenters to distribute their research, thereby benefiting typological researchers studying cross-linguistic differences in language processing.

Participants remain fully anonymous, identified internally only by their IP address to prevent them from retaking the same experiment. Participants can also decide whether to authorize access to their experimental data to researchers other than the original experimenters in the online consent form. Data access authorized only for the original experimenter remains hidden to all other users even when the experimenter shares the results in Worldlikeness. Participants can also choose to withdraw from an experiment at any time simply by closing the Web browser, and no experimental data will be stored at all.

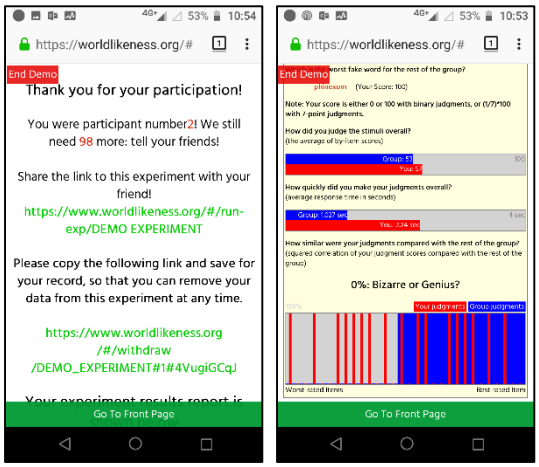

Figure 3. Experiment-sharing and data-withdrawing links (left) and participant performance (right) on the result report page in mobile Mozilla Firefox v64 in Android 8.0

At the end of an experiment, participants may be invited to give feedback on the clarity of the consent form and instructions and how likely they would recommend this experiment to others. All participants then receive a personalized results report (Figure 3) that compares their mean response rates and reaction times to the rest of the participants in the same experiment. The feedback page also includes a direct link to the experiment so that participants can help recruit further participants. Another link allows participants to withdraw their data from the Worldlikeness server at any time in the future.

Researchers can then use the desktop interface to search for and download the experimental materials and results that experimenters and participants have agreed to share, as illustrated in Figure 4. Researchers who set up experimenter accounts may gain access to more data.

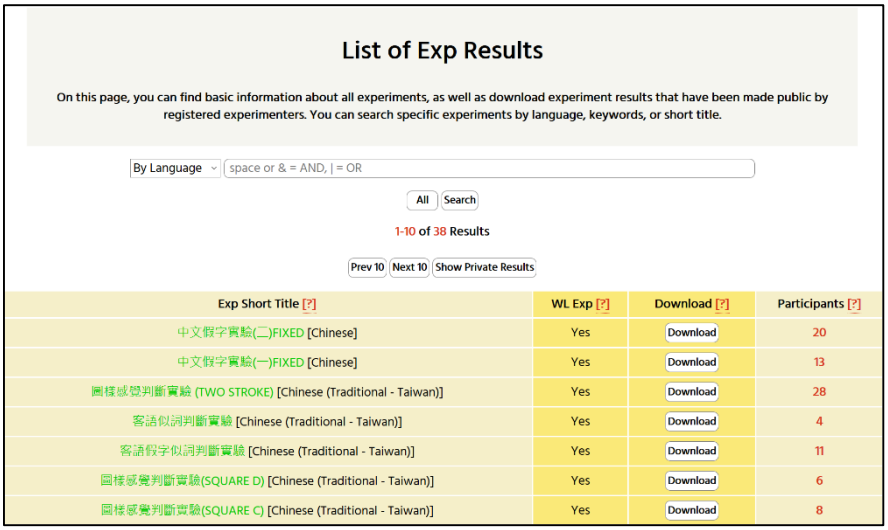

Figure 4. Publicly shared results for experiments tagged with “Chinese” available through the desktop interface

In short, Worldlikeness is designed not only to facilitate psycholinguistic crowdsourcing from participants using either desktop or mobile devices, but also to foster the development of a cross-linguistic database by offering incentives to share data and by simplifying and standardizing the process as much as possible, at least for wordlikeness studies, while protecting experimenter and participant rights.

3 Reliability of wordlikeness judgment data in Worldlikeness

To test the validity and linguistic value of Worldlikeness, we compared results from in-lab and online participants in two wordlikeness judgment studies.

3.1 Replication of Mandarin wordlikeness megastudy

Our first test was a small-scale Mandarin wordlikeness judgment experiment run in Worldlikeness that aimed to replicate a megastudy run using E-Prime (Schneider et al. 2002). In the original megastudy, reported in Myers (2015), over 3,000 nonwords were divided into two random item sets (blocks) and presented in Zhuyin Fuhao (the onset-rime-based phonetic orthography used in Taiwan) to more than 100 native Mandarin speakers. Participants were asked to provide binary judgments on whether each nonword was like Mandarin or not (i.e., acceptance as Mandarin-like). A key finding was that a nonword was significantly more likely to be judged as being like Mandarin the more phonological neighbors it had in the Mandarin lexicon (differing in only one phoneme, ignoring tone), similar to results in other languages (e.g., Bailey and Hahn 2001, for English).

To replicate this finding in Worldlikeness, we randomly selected four practice items and 100 test items from each of the two nonword sets used in the megastudy, and created two respective experiments in Worldlikeness, which were run online via the crowdsourcing process. A link to the experiments was distributed via e-mail to undergraduate students and advertisements posted on Facebook, starting in Jan 24, 2019, specifically encouraging them to use a smartphone. Once directed to Worldlikeness by the link, participants were randomly assigned to one of the two experiments automatically using the experiment grouping feature and asked to provide binary

judgments ('like Mandarin' vs. 'unlike Mandarin'), as in the megastudy.[2] A total of 43 participants completed the experiments (N = 22 for Set 1 and N = 21 for Set 2) by Feb 3, 2019, with self-reported age 20-47 (mean = 27.1, sd = 7.2). We excluded 6 participants (N = 4 for Set 1 and N = 2 for Set 2) who were identified as possible users of a desktop device (with a screen resolution wider than 1,000 pixels). Among the rest of the mobile users, 21 participated with the mobile Safari browser on a device running Apple's iOS, and 16 completed an experimental session with a mobile Chrome browser on an Android device.

The distributions of reaction times excluding practice trials, no response trials (N = 24 [0.3%] in our replication and N = 24 [0.1%] in the megastudy), and trials responded within 100 ms (N = 0 in our replication and N = 853 [3.9%] in the megastudy) in Figure 5 suggest a high similarity in the time data collected from mobile users using Worldlikeness and E-Prime, regardless of the experiment.

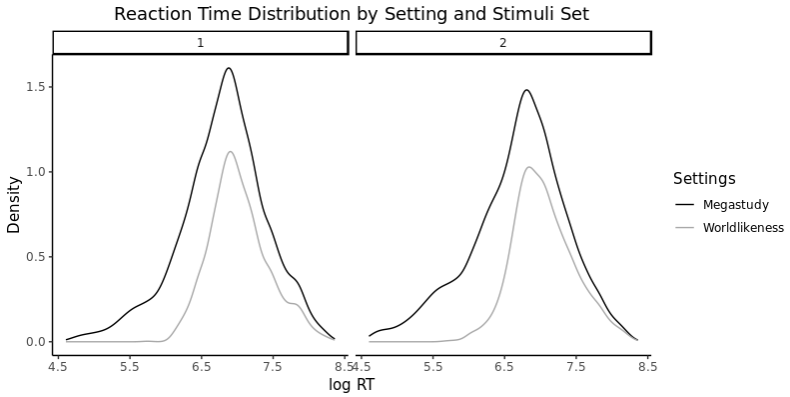

Figure 5. Distribution of log-transformed reaction times (RT) in the original megastudy and in the Worldlikeness replication

By-item judgment scores and reaction times from the above data set were then compared across the two studies in a linear regression model. For each item in our replication and the megastudy, judgment scores and log reaction times were averaged across participants. Within each item set, mean judgment scores and mean log reaction times were z-scored. In the linear regression models, mean judgment and log reaction time z-scores from the megastudy served as the sole predictors of their counterparts in our replication within each item set. As illustrated in Figure 6, the z-scored mean acceptance rates and reaction times from the megastudy was a significant predictor of those from our replication in both item sets (Set 1 Judgment: β = 0.685 SE = 0.074, t = 9.31, p < .001; Set 2 Judgment: β = 0.62, SE = 0.792, t = 7.83, p < .001; Set 1 RT: β = 0.433, SE = 0.091, t = 4.76, p < .001; Set 2 RT: β = 0.449, SE = 0.09, t = 4.97, p < .001).

[2] The two data sets are available at https://www.worldlikeness.org/#/resultsInfo/mKKFSmNXDgsfwjpdP (Set 1) and https://www.worldlikeness.org/#/resultsInfo/zRa8eT5wHMgEcvBLf (Set 2). The numbers of participants available through these public links differ from those reported here as some participants chose not to authorize their data for public use.

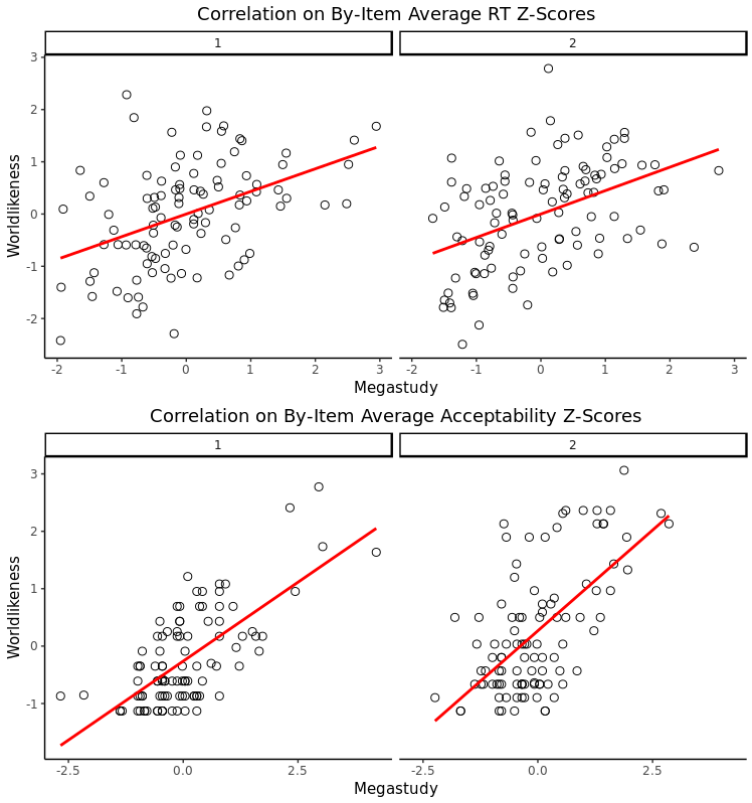

Figure 6. By-item comparisons of log-normed reaction times (RT) and mean acceptance rates in the Worldlikeness replication with the original megastudy for the two stimuli sets

The experimental results were then analyzed using mixed-effect logistic regression, with wordlikeness judgments as the dependent variable, log-transformed neighborhood density *z*-scores as the sole independent variable, random intercepts for participant and items, and by-participant random slopes for neighborhood density. Unsurprisingly, we observed the same positive neighborhood density effect for both item sets (Set 1: $\beta = 1.73$, $SE = 0.33$, $z = 5.23$, $p < .001$; Set 2: $\beta = 1.69$, $SE = 0.24$, $z = 7.09$, $p < .001$): more lexical neighbors lead to higher acceptability. These strong correlations and the replication of the neighborhood density effect suggest that judgment data collected from smartphone users via Worldlikeness may be as reliable as those collected in-lab using E-Prime.

3.2 Cross-linguistic judgments of auditory stimuli

To explore how Worldlikeness can contribute to typological psycholinguistics, we conducted a study examining wordlikeness judgments for an identical set of nonword stimuli by bilingual speakers of two related but crucially distinct languages. Again, we also compared participants tested in-lab and online (the mobile-specific interface was not yet available when we ran this earlier experiment, though as a Web-based system, Worldlikeness has always been accessible via any internet-connected device).

Our linguistic interest concerned lexical influences on the acceptability of monosyllabic nonwords presented auditorily to bilingual speakers of Mandarin and Taiwan Southern Min (commonly called Taiwanese). While these two major languages spoken in Taiwan are members

of the Sinitic language family, they differ considerably in their phonological systems; in particular, Mandarin has a simpler syllable structure, resulting in Southern Min having almost twice as many lexical syllables. Moreover, Mandarin not only has a logographic orthography but also the Zhuyin Fuhao phonetic writing system, actively used for typing, whereas Southern Min is still virtually never written, despite official writing systems being promulgated since the 1990s.

Our experimental materials were 200 syllables that are nonwords in both Mandarin and Southern Min, randomly selected from all possible combination of onsets, rhymes, and lexical tones in Mandarin and Southern Min.[3] These were written in IPA and read aloud by two female speakers whose home languages were, respectively, Mandarin (speaker KY) and Southern Min (speaker PS).[4] A subset of 71 items were excluded for being judged by twelve Mandarin-Southern Min bilinguals as too confusable with real words either in Mandarin or Southern Min.

We recruited 81 in-lab bilingual speakers of Mandarin and Southern Min (55 males and 26 females), aged 18 to 28 years (mean = 21.6, *sd* = 1.95). Another 156 online bilingual speakers (103 males, 52 females, 1 transgender), 18 to 62 years (mean = 26.8, *sd* = 7.6), were recruited in less than two weeks via a Facebook advertisement.[5] All participants were randomly assigned to judge items either for Mandarin or Southern Min wordlikeness, with items produced by either speaker (2 Speaker × 2 Target Languages). Both in-lab and online participants had to pass a proficiency test in the assigned language prior to beginning the experiment itself. The participants expressed their binary judgments of 'unlike' or 'like' Mandarin by respectively pressing 'S' or 'L' on the keyboard or by tapping screen buttons labeled 'Unlike (S)' or 'Like (L)'.

For illustrative purposes, our analysis and discussion below only include a subset of the results from experimental sessions where the target language and the home language of the two speakers were consistent.[6] Participants under 20 years old were also excluded in accordance with local regulations. This left the Mandarin subset with 17 in-lab participants (age = 20-26, mean = 22.5, *sd* = 1.8) and 35 online participants (age = 20-43, mean = 27.5, *sd* = 5.7), and the Southern Min subset with 19 in-lab participants (age = 20-26, mean = 21.2, *sd* = 1.5) and 35 online participants (age = 20-54, mean = 28, *sd* = 7.6).

This data set, excluding trials with no or very short responses (< 100 ms), was submitted to a mixed-effect logistic regression model, in which the dependent variable was binary wordlikeness judgments (acceptability) and the independent variables were target language (Mandarin vs. Southern Min), log-normed target language neighborhood density (ignoring tones),

[3] These nonwords did not include any mid-level tone and obstruent codas, which were only possible in Southern Min and would therefore violate Mandarin phonotactics too blatantly.

[4] The speakers' different accents also allowed us to test the effects of sociophonetic factors on wordlikeness judgments, but for space considerations we do not discuss these here (though see Myers and Chen 2016).

[5] The Facebook post https://www.facebook.com/lngproc.exp/posts/1355912174498901 has been viewed 13,232 times as of Jan 25, 2018.

[6] See Myers and Chen (2016) for more details. The raw results are also available at https://www.worldlikeness.org/#/resultsInfo/yACHq5Et9fksc6Mfs (In-lab Mandarin I), https://www.worldlikeness.org/#/resultsInfo/TvmjnxSuNpTpBuG5M (In-lab Mandarin II), https://www.worldlikeness.org/#/resultsInfo/Qdr8LpmB6ym4DiLAp (In-lab Southern Min I), https://www.worldlikeness.org/#/resultsInfo/No2ATWirJPoeN5ZDZ (In-lab Southern Min II), https://www.worldlikeness.org/#/resultsInfo/TqqvJGeZ6wi99x3hq (Online Mandarin I), https://www.worldlikeness.org/#/resultsInfo/hvr7cRFsLeEEb9Yh4 (Online Mandarin II), https://www.worldlikeness.org/#/resultsInfo/3BYBHaPttKu3EeAzN (Online Southern Min I), and https://www.worldlikeness.org/#/resultsInfo/JA8n3i2QvCCb439g6 (Online Southern Min II). The number of participants accessible through these public links are different from those we report here since not all participants chose to publicly authorize their data.

setting (in-lab vs. online), and all interactions, along with random intercepts for participants and items (the model was too complex to include random slopes). The results for acceptability judgments are illustrated in Figure 7. Acceptability was significantly lower for Mandarin targets than for Southern Min targets (β = -0.67, *SE* = 0.22, *z* = -2.99, *p* = .003), as well as for judgments made in the online crowdsourcing context (β = -0.49, *SE* = 0.18, *z* = -2.7, *p* = .007). As expected, items with greater neighborhood density were more likely to be accepted (β = 0.56, *SE* = 0.06, *z* = 9.3, *p* < .001), but this effect was significantly weaker for Mandarin (β = -0.22, *SE* = 0.1, *z* = -2.3, *p* = .023) and in the online crowdsourcing context (β = -0.14, *SE* = 0.06, *z* = -2.4, *p* = .016). There was no significant three-way interaction (*p* > .14), though the graphs show a trend for test setting to matter more for Southern Min than for Mandarin. This trend suggests that despite our language proficiency pretest, online participants may be particularly hard to vet for fluency in a language that is not only unwritten, but socially less prestigious (Chen 2010). Potential distractions faced by online participants, such as multitasking, ambient noise, and variations in volume of the participants' devices, may have further exacerbated the difficulty that participants may already face with processing Southern Min auditorily online.

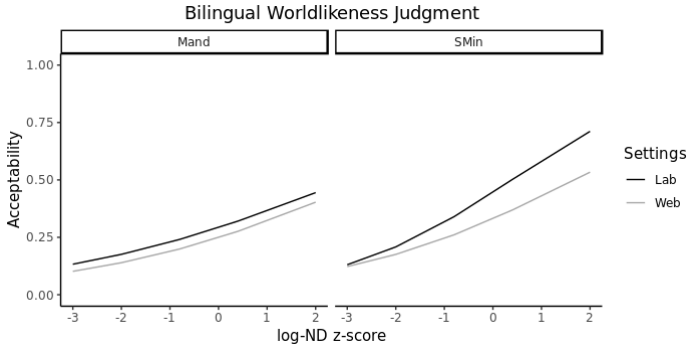

Figure 7. Bilingual wordlikeness judgment by Target Language, Target Language neighborhood density (ND), and Setting

In addition to the methodological implications of this study, the significant interaction between language and neighborhood density is also theoretically intriguing. Since neighborhood density effects are defined on whole syllables, they may indicate a lexical access path where speakers do not decompose syllables into phonemes. Our results thus suggest that Southern Min speakers follow this path more than Mandarin speakers. As noted above, previous experiments have shown that English speakers activate phonemes more readily than Mandarin speakers, but it is not clear if this is because of differences in orthography or in syllable inventory sizes (the smaller number of lexical syllables in Mandarin may encourage whole-syllable access). Since, as also noted above, Southern Min has a larger syllable inventory than Mandarin but no widely used phonetic writing system, our results hint that the crucial cross-linguistic factor affecting syllable decomposition may be orthography, not syllable inventory size. This interpretation is not conclusive, of course, since we compared only two languages, too few to take other cross-linguistic factors, let alone sampling bias, into account. In an ongoing Worldlikeness meta-megastudy (see below), however, we are collecting and comparing a larger number and greater variety of languages to tease apart the typological variables affecting syllable and phoneme processing.

4 Future challenges

In this paper, we have highlighted the importance of studying typological psycholinguistics via Web crowdsourcing, and showed how Worldlikeness was developed to help conduct simple and ethically grounded studies on native speakers of different languages, using both desktop and mobile devices. In addition to the Mandarin and Southern Min data we have discussed, the database now includes the other Sinitic languages Cantonese and Hakka, and all four were recently included in a study of micro-variation in syllable decomposition by listeners as a function of a language's syllable inventory size (Myers and Chen 2019). We have also completed pilot wordlikeness studies on the non-Sinitic languages Indonesian and Vietnamese, and are currently working on collecting and sharing wordlikeness data from English, Japanese, and Polish, to increase the typological diversity of the online database. While we will conduct cross-linguistic analyses ourselves on all of this data, our primary goal is to "seed" the Worldlikeness database in the hope that other researchers will be inspired to make further contributions of their own, both to study particular languages of interest and to conduct cross-linguistic studies larger than any single research group would be capable of on its own.

Other minor but essential improvements continue to be made. On the technical side, our priorities are to further reduce the amount of data that needs to be downloaded to the clients' browser and to keep pace with rapid changes in HTML5, JavaScript, and individual browsers. Meanwhile, we are also working to provide more flexible experimental paradigms in Worldlikeness, making minor updates to allow researchers to investigate different aspects in language processing cross-linguistically. For example, researchers will soon be able to insert a frame for a within-trial prime, so that combined with the system's existing ability to handle many types of stimuli, cross-modal priming experiments will soon be possible within Worldlikeness. We also plan to modify the Web interface so that it can accommodate the self-paced moving window paradigm commonly adopted in the study of sentence processing (as in the Linger program; Rohde 2003). In additional to these technical improvements, we are also working to solve human ones, such as the problem of motivating online participants; in our megastudy replication, only around one fifth of the online participants who initiated an experimental session eventually completed it. One initial solution is to distribute online recruitment messages more effectively via social networking sites to a larger number of internet users, for a larger potential subject pool. A more fundamental solution, perhaps, would be to make Worldlikeness more rewarding by enhancing its gamification elements (see e.g., Harami et al. 2014), although this may require participants to sign up for an account to allow them to keep track of their individual rewards, a shift away from the current total-anonymity plan of Worldlikeness. Moreover, administrative requirements have required Worldlikeness to present lengthy and detailed consent forms prior to every experimental session, which may further discourage online participation. We have already streamlined this requirement on the technical side, making the participant experience similar to dealing with the end-user license agreement click-to-agree screens ubiquitous in the software world, but ultimately this issue is a matter for local ethics review boards, whose standards and principles may vary across the world. For online megastudy and meta-megastudy systems to reach their full potential, as with digital technology more generally, a balance has to be struck between convenience and ethics.

Acknowledgment: This study was funded by the Ministry of Science and Technology, Taiwan (103-2410-H-194-119-MY3; 106-2410-H-194-055-MY3; 107-2410-H-007-002-MY2) and was approved by the Research Ethics Committees at National Cheng Kung University and National Chung Cheng University (CCU). We are particularly grateful to the CCU committee for

collaborating with us on accommodating online studies within their guidelines. We also thank the following lab assistants who have helped test and improve Worldlikeness, and set up and run in-lab and online experiments: Guan-Nan Chiang, Jia-Cing Ruan, Kuei Mei-Chun Liu, Pei-Shan Chen, Shiao-Yin Pan, Si-Qi Su, Yan-Jhe Ciou, Yi-Hsin Wu, Yu-Chu Chang, and Yu-Shan Yen. The editors and two anonymous reviewers are greatly appreciated for their valuable comments on our work. All errors remain our own.

References

Aker, Jenny C. & Isaac M. Mbiti. 2010. Mobile phones and economic development in Africa. *Journal of Economic Perspectives* 24(3). 207–232.

Antoun, Christopher, Mick P. Couper & Frederick G. Conrad. 2017. Effects of mobile versus PC web on survey response quality. *Public Opinion Quarterly* 81(S1). 280–306.

Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44(4). 568–591.

Balota, David A., Melvin J. Yap, Keith T. Hutchison & Michael J. Cortese. 2012. Megastudies: What do millions (or so) of trials tell us about lexical processing? In James S. Adelman (ed.), *Visual Word Recognition, Vol. 1: Models and methods, orthography, and phonology*, 90–115. London, UK: Psychology Press.

Chen, Jenn-Yeu, Train-Min Chen & Gary S. Dell. 2002. Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language* 46(4). 751–781.

Chen, Su-chiao. 2010. Multilingualism in Taiwan. *International Journal of the Sociology of Language*, 205. 79–104.

Dryer, Matthew S. & Martin Haspelmath. 2013. *The world atlas of language structures*. https://wals.info/. (accessed 6 February 2018).

Enochson, Kelly & Jennifer Culbertson. 2015. Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PLOS ONE* 10(3). e0116946.

Frank, Michael C., Elika Bergelson, Christina Bergmann, Alejandrina Cristia, Caroline Floccia, Judit Gervain, J. Kiley Hamlin, et al. 2017. A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy* 22(4). 421–435.

Frisch, Stefan A., Nathan R. Large & David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42(4). 481–496.

Hamari, Juho, Jonna Koivisto & Harri Sarsa. 2014. Does gamification work? – A literature review of empirical studies on gamification. *The proceedings of the 47th Hawaii International Conference on System Sciences*. 3025–3034.

Hayes, Bruce & James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44(1). 45–75.

Keller, Frank, Subahshini Gunasekharan, Neil Mayo & Martin Corley. 2009. Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods* 41(1). 1–12.

Keuleers, Emmanuel, Michaël Stevens, Paweł Mandera & Marc Brysbaert. 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.

Lee, Hsin-Hsien 2016. *A comparative study of the phonology of Taiwan Sign Language and Signed Chinese*. Ph.D. thesis, National Chung Cheng University.

McCarthy, John J. & Prince, Alan. 1996. Prosodic morphology 1986. Linguistics Department Faculty Publication Series. 13. https://scholarworks.umass.edu/linguist_faculty_pubs/13. (accessed 6 February 2018).
Myers, James. 2015. Markedness and lexical typicality in Mandarin acceptability judgments. *Language and Linguistics* 16(6). 791–818.
Myers. James. 2016. Meta-megastudies. *The Mental Lexicon* 11(3). 329–349.
Myers, James. 2019. *The grammar of Chinese characters: Productive knowledge of formal patterns in an orthographic system*. London: Routledge.
Myers, James & Tsung-Ying Chen. 2016. The time course of sociolinguistic influences on wordlikeness judgments. *The proceedings of ExLing 2016*. 119–122.
Myers, James & Tsung-Ying Chen. 2019. Variation in syllable decomposition across Sinitic languages. Paper presented at The 27th Annual Conference of International Association of Chinese Linguistics, Kobe City University of Foreign Studies, 10-12 May.
Norcliffe, Elisabeth, Alice C. Harris, & T. Florian Jaeger. 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, Padraig G., Jenn-Yeu Chen & Train-Min Chen. 2010. Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition* 115(2). 282–302.
Paolacci, Gabriele, Jesse Chandler & Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5(5). 411–419.
Peirce, Jonathan W. & Michael R. MacAskill 2018. *Building experiments in PsychoPy*. London: Sage.
Rohde, Doug. 2003. *Linger: a flexible platform for language processing experiments*. https://tedlab.mit.edu/~dr/Linger/. (accessed 26 Jun 2019)
Ryan, Kenneth John, Joseph V. Brady, Robert E. Cooke, Dorothy I. Height, Albert R. Jonsen, Patricia King, Karen Lebacqz, David W. Louisell, Donald W. Seldin, Eliot Stellar & Robert H. Turtle. (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, Walter, Amy Eschman & Anthony Zuccolotto. (2002). *E-Prime: User's guide*. Pittsburgh, PA: Psychology Software Incorporated.
Snyder, William. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31(3). 575–582.
Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1). 155–167.
Tsay, Jane, James H.-Y. Tai, & Yijun Chen. 2017. Taiwan Sign Language online dictionary, 3rd English edition. Institute of Linguistics, National Chung Cheng University, Taiwan. http://tsl.ccu.edu.tw/web/browser.htm. (accessed 6 February 2018).
von Bastian, Claudia, André Locher & Michael Ruflin. 2013. Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods* 45(1). 108–115.