Complexity of Chinese characters (Oracle vs Traditional) (#66462)

Created: 05/20/2021 09:54 PM (PT)
Public: 09/13/2021 03:13 AM (PT)

Author(s)
Charles Kemp (University of Melbourne) - c.kemp@unimelb.edu.au
Jerome Han (University of Melbourne) - simon.jerome.han@gmail.com
Piers Kelly (University of New England) - pkelly26@une.edu.au
James Winters (University Mohammed VI Polytechnic) - james.winters@um6p.ma

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
That traditional Chinese characters are perceived as more complex than their counterparts from the Oracle Bones period.

3) Describe the key dependent variable(s) specifying how they will be measured.
Participants will be shown pairs of characters, and asked to judge which member of the pair is more complex. Ratings will be provided on a 6 point scale:

L3: left character much more complex than the right character
L2: left character moderately more complex than the right character
L1: left character slightly more complex than the right character

and R1, R2, and R3 are defined analogously.

4) How many and which conditions will participants be assigned to?
Two between-participant conditions. In the "printed" condition participants will compare Oracle characters with traditional characters shown in Hiragino Sans GB font. In the "handwritten" condition traditional characters will be handwritten forms drawn from the traditional Chinese handwriting dataset available at:

https://github.com/AI-FREE-Team/Traditional-Chinese-Handwriting-Dataset

This data set includes multiple versions of each character, and we will present the version with median perimetric complexity.

Both conditions will use all characters that are (1) classified as "pictographic" in the Chinese Lexical Database; (2) have at least one Oracle Bones form available from hanziyuan.com; and (3) are included in the Traditional Handwriting Dataset linked above. There are 155 such characters. Each participant will make judgments about 50 characters randomly drawn from the full set.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Each response will be converted into a binary rating r (0 = oracle version more complex, 1 = traditional version more complex). These ratings will then be analyzed using a mixed-effects logistic regression with random effects for participant and character:

M1: r ~ 1 + (1|participant) + (1|character)

For each condition, we will carry out

(i) a Bayesian analysis that uses brms to compute the posterior distribution on the intercept of M1 (hypothesis is supported if the 95% credible interval excludes 0 and has a mean greater than 1)

(ii) a frequentist analysis that uses a likelihood ratio test to compare M1 with a null model that has zero intercept:

M0: r ~ 0 + (1|participant) + (1|character)

(hypothesis supported if the M1 intercept exceeds zero and if the model comparison supports M1 rather than M0 with $p < 0.05$)

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Each participant will judge 5 "check" pairs in addition to 50 "real" pairs. Each check pair (c1, c2) has a c1 that is a proper subset of c2, so c2 is unambiguously more complex than c1. Participants who make any error on these check pairs will be excluded.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

For each condition we'll aim to have roughly 40 judgments about each of 155 characters (40 * 155 = 6200 judgments in total). Each participant will provide 50 judgments, which means that we'll need at least 124 participants. To allow for exclusions we'll recruit 200 participants in each condition.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

If our hypothesis is supported, we'll repeat the analysis including only the 125 characters for which the traditional form matches the simplified form. The results of this follow up analysis provide information about whether simplified forms tend to be more complex than their Oracle counterparts.

We'll also use the data to discuss whether the printed and handwritten conditions lead to different patterns of results, but this comparison is not of primary interest.

At the end of the survey we'll ask people to rate how much they know about Chinese characters, but will not use responses to this question in our main analyses.