Emmy Shi PSY503 Final Project: Re-Analysis of Cue Validity Effects Using Open Behavioral Data

Yujingai (Emmy) Shi[1]

[1] Princeton University

Author Note

Correspondence concerning this article should be addressed to Yujingai (Emmy) Shi, Princeton University, Department of Psychology. E-mail: es2277@princeton.edu

Abstract

This project re-analyzes open behavioral data from Schmitz et al. (2024) to examine whether cue validity influences reaction time using statistical methods learned in PSY503 such as ANOVA test, Regression test, and power analysis. Focusing on Experiment 1 data from the original paper, I reproduced key analyses comparing valid versus invalid cues and differences between arrow and gaze cues. The results revealed small validity effects and no significant interaction with cue type since the data size is limited and the design of the experiment leads to very simple performances. A simulation-based power analysis showed that very large samples would be required to reliably detect this effect.

Keywords: statistics, eyetracking, attention tests, psychophysics

Emmy Shi PSY503 Final Project: Re-Analysis of Cue Validity Effects Using Open Behavioral Data

Introduction

Selective attention is a fundamental skills for helping people direct their attention to various positions in a dynamic and complex environment. In laboratory settings, Attentional cueing paradigms are widely used to study how people orient their attention in visual environments. There are many classic findings by using attentional cueing paradigms or gabor patch paradigms show that people's responses are faster when a cue correctly predicts the location of a target; This is an example of demonstrating a robust "validity effect."

The dataset analyzed in this project comes from Experiment 1 of this openly available study comparing gaze cues and arrow cues. I only selected Experiment 1 data since there are various different experiments conducted in this study; these experiments are layered in a complex form, so I only selected experiment 1 to show a clean and straightforward analysis. The goal of the present analysis was not to fully replicate the original paper, but instead to reproduce core analyses to examine attention effects, examine basic patterns in reaction time data, and estimate the statistical power required to detect the observed effects. This analysis provided an opportunity for me to apply the statistical tools learned in class to real behavioral data, and through these analysis, I have developed clearer understanding on a study's robustness and power.

Method

Below I am listing the Experiment 1 set up and procedure, and the data analysis pipeline I used to examine the attentional effects and the power tests. ## Procedure Experiment 1 used a spatial cueing paradigm; each trial began with either an eye-gaze cue (a face whose eyes shifted left or right) or an arrow cue pointing in one direction. After a brief cue–target interval, a target letter appeared on either the left or right side of the display, and participants responded by pressing a key to record. On valid trials the cue correctly indicated the target location; on invalid trials it pointed in the opposite direction. Reaction time and accuracy were recorded on every trial. All data for Experiment 1 were publicly available through the OSF repository associated with the original article. ## Data analysis In this project, I re-analyzed Experiment 1 using the statistical tools learned in PSY503. The primary analyses focused on computing the classic validity effect (valid vs. invalid trials) and testing whether this effect differed between gaze cues and arrow cues through ANOVA and Regression tests. I also conducted supplementary analyses, including visualization of reaction time and accuracy performance, and a simulation-based power analysis to assess the sample size required to detect the observed effect sizes.

Results

The dataset is saved in the project repo under **data/Exp1_RT.RDS**. Using a **relative path** ensures complete reproducibility.

```
## Warning: package ’ggplot2’ was built under R version 4.5.2
```

```
##   VP block trial cueType     cue tgt tgtLoc resp   RT valid train valid2
## 1  1     1     1    face   right   2  right    A 1218  TRUE train   TRUE
## 2  1     1     2    face    left   1  right    B  656 FALSE train  FALSE
## 3  1     1     3    face    left   2  right    A  445 FALSE train  FALSE
## 4  1     1     4    face neutral   2  right    A  371 FALSE train     NA
## 5  1     1     5    face   right   1   left    B  511 FALSE train  FALSE
## 6  1     1     6    face   right   2   left    A  527 FALSE train  FALSE
##   instruction tgtResp correct     cueDir     tgtDir
## 1           A       2    TRUE horizontal horizontal
## 2           A       1    TRUE horizontal horizontal
## 3           A       2    TRUE horizontal horizontal
## 4           A       2    TRUE    neutral horizontal
## 5           A       1    TRUE horizontal horizontal
## 6           A       2    TRUE horizontal horizontal
```

Next, I followed standard preprocessing used in attentional cueing experiments: 1. keep only test trials 2. remove extreme reaction times (<150 ms or >2000 ms) 3. use valid2 (TRUE/FALSE) as the correct validity coding 4. convert categorical variables to factors

And we can see that we have 3830 valid trials where cue predicted the target and 3829 invalid trials where cue misled the target, so this means the project has balanced trials, which is good for further analysis.

```
## Rows: 11,495
## Columns: 17
## $ VP          <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ block       <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ trial       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ cueType     <fct> face, face, face, face, face, face, face, face, face, face~
## $ cue         <chr> "left", "neutral", "neutral", "right", "neutral", "right",~
## $ tgt         <dbl> 1, 2, 1, 1, 2, 2, 2, 2, 2, 1, 1, 2, 1, 1, 2, 2, 1, 2, 2, 2~
## $ tgtLoc      <chr> "right", "right", "right", "left", "left", "left", "right"~
## $ resp        <chr> "B", "A", "B", "B", "A", "A", "A", "A", "A", "B", "B", "A"~
## $ RT          <dbl> 618, 442, 412, 548, 445, 413, 418, 369, 374, 354, 333, 365~
## $ valid       <fct> invalid, NA, NA, invalid, NA, invalid, invalid, invalid, v~
## $ train       <chr> "test", "test", "test", "test", "test", "test", "test", "t~
## $ valid2      <lgl> FALSE, NA, NA, FALSE, NA, FALSE, FALSE, FALSE, TRUE, NA, F~
## $ instruction <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A"~
## $ tgtResp     <dbl> 1, 2, 1, 1, 2, 2, 2, 2, 2, 1, 1, 2, 1, 1, 2, 2, 1, 2, 2, 2~
## $ correct     <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE~
## $ cueDir      <chr> "horizontal", "neutral", "neutral", "horizontal", "neutral~
## $ tgtDir      <chr> "horizontal", "horizontal", "horizontal", "horizontal", "h~
```

```
## 
##   valid invalid 
##    3830    3829
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   179.0   370.0   432.0   460.5   510.0  1983.0
```

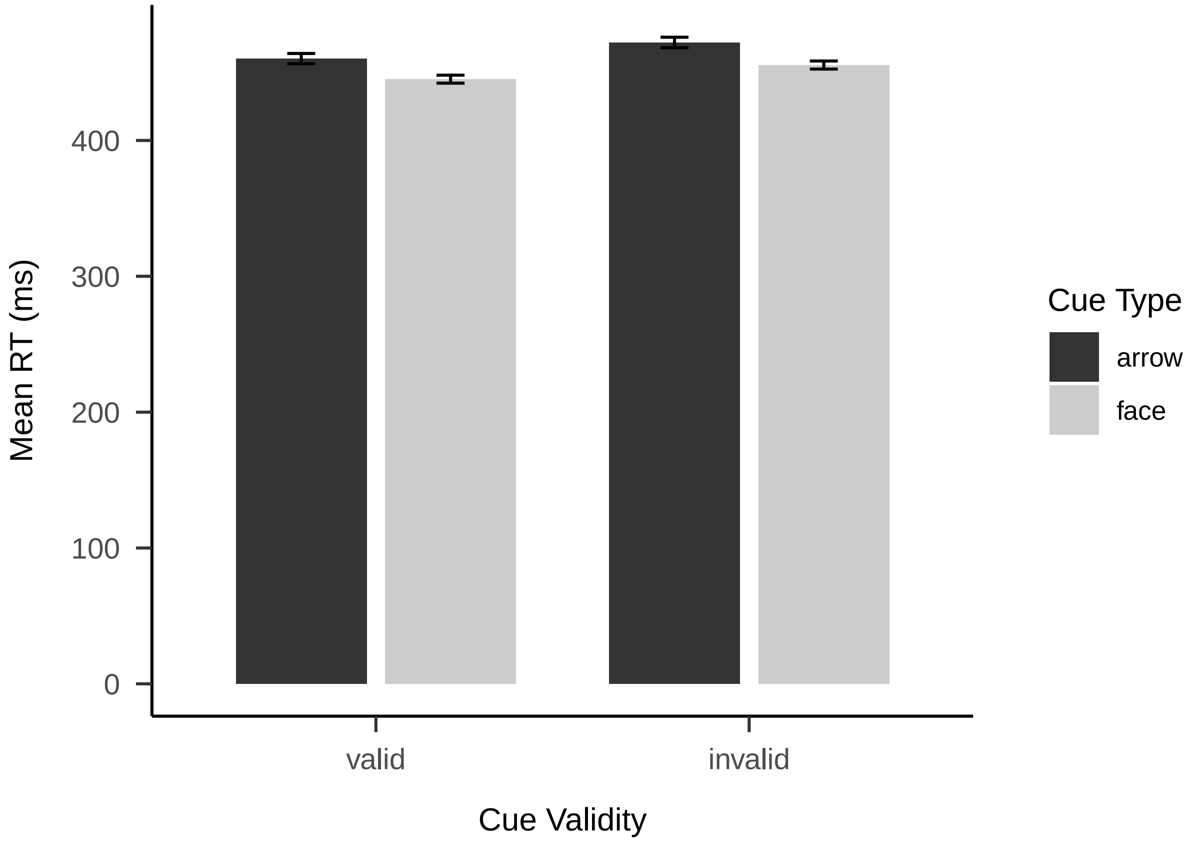
Descriptive Statistics: to compute descriptive statistics by cuetype and validity

Before the main analyses, we could compute **summary statistics** for each condition (cue type × validity). This helps check whether the expected validity effect appears in both cue types. For each condition,I computed the mean RT(reaction time), standard deviation, sample size, and the standard error of the mean (SEM).

```
## # A tibble: 4 x 6
##   cueType valid   mean_RT sd_RT     n se_RT
##   <fct>   <fct>     <dbl> <dbl> <int> <dbl>
## 1 arrow   valid      460.  166.  1913  3.79
## 2 arrow   invalid    472.  168.  1917  3.85
## 3 face    valid      445.  129.  1917  2.94
## 4 face    invalid    456.  130.  1912  2.97
```

The table displays the mean reaction time, standard deviation, sample size (n), and standard error (SE) for each combination of cue type (arrow vs face) and cue validity (valid vs invalid). **We could tell that people recognize face cues are overall faster than arrow cues, which is normal here**

Below is the visualization for the descriptive data



This plot shows mean reaction times for valid versus invalid cues,for arrow cues and face(gaze) cues.Face cues produce overall faster reaction times than arrow cues, and valid trials are faster than invalid trials.

2 *2 ANOVA on Reaction Time Now I want to specifically know that 1)Are reaction times faster on valid versus invalid trials?, and 2) Does the size of the validity effect differ for arrow versus face cues? So I have a 2x2 ANOVA. I fit a two-way ANOVA predicting RT from cue type (arrow vsface), cue validity (valid vs invalid), and their interaction.Only trials with clear validity labels (valid/invalid) are included here.

```
##                   Df    Sum Sq Mean Sq F value   Pr(>F)    
## cueType            1    483078  483078  21.647 3.33e-06 ***
## valid              1    237833  237833  10.658   0.0011 ** 
## cueType:valid      1      1039    1039   0.047   0.8292    
## Residuals     7655 170828875   22316
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 3836 observations deleted due to missingness
```

From the ANOVA test, there is a significant main effect of cue type (F= 21.647, p < .001); responses to face cues were slightly faster than to arrow cues and this is statistically significant. In addition, there is a significant main effect of validity (F= 10.66, p = .0011); Participants responded faster on valid trials than invalid trials, **replicating the classic attentional validity effect**. Finally, there is a non-significant cueType and validity interaction

Effect Size I also calculated the eta squred to reflect the proportion of variance in RT explained by each effect while controlling for others. This is just a complementary measure here.

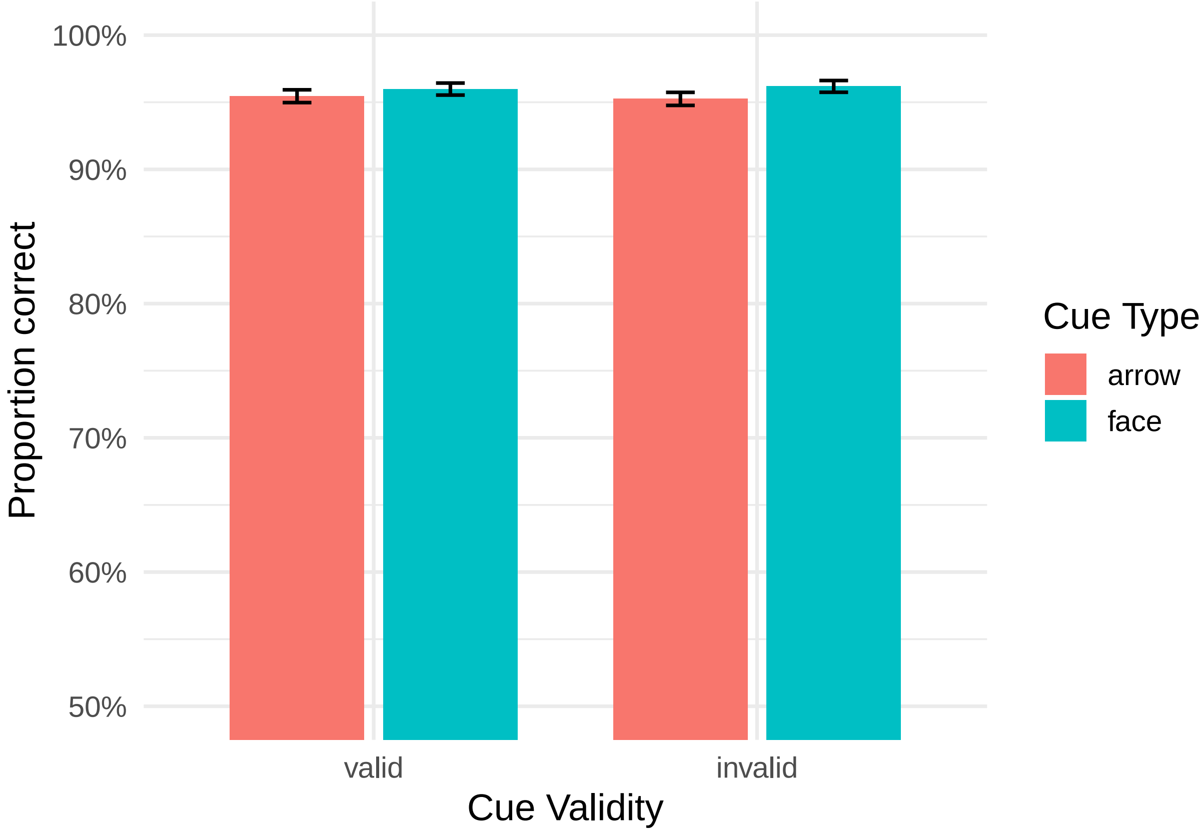
```
## # Effect Size for ANOVA (Type I)
##
## Parameter     | Eta2 (partial) |       95% CI
## ---------------------------------------------
## cueType       |       2.82e-03 | [0.00, 1.00]
## valid         |       1.39e-03 | [0.00, 1.00]
## cueType:valid |       6.08e-06 | [0.00, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].
```

Although the partial eta-squared values are small (< .01), this pattern is typical for reaction time data, where variability is influenced by many cognitive and motor processes.I think that small effect sizes do not indicate a problem with the analysis;rather, they reflect that the experimental manipulations account for a modest portion of RT variance. Finally, this is consistent with what is commonly observed in attention research.

Now I have calculated the accuray for the different cue types and validty status. Next, I am going to perform analysis on accuracy.

```
## # A tibble: 4 x 5
##   cueType valid   mean_acc     n  se_acc
##   <fct>   <fct>      <dbl> <int>   <dbl>
## 1 arrow   valid      0.955  1913 0.00476
## 2 arrow   invalid    0.953  1917 0.00486
## 3 face    valid      0.960  1917 0.00448
## 4 face    invalid    0.962  1912 0.00438
```

I calculated **accuracy performance across cue types and cue validity**. The purpose of this analysis is to confirm that participants performed the task well overall. I computed the proportion of correct responses for each combination of cueType (arrow vs. face) and valid (valid vs. invalid), along with standard errors.The following plot shows those information.



Accuracy was **uniformly very high** across all conditions (approximately 95–97%), with very small differences between cue types or validity conditions. Importantly, no systematic drop in accuracy was observed for invalid trials, suggesting that participants did not sacrifice accuracy to respond more quickly.

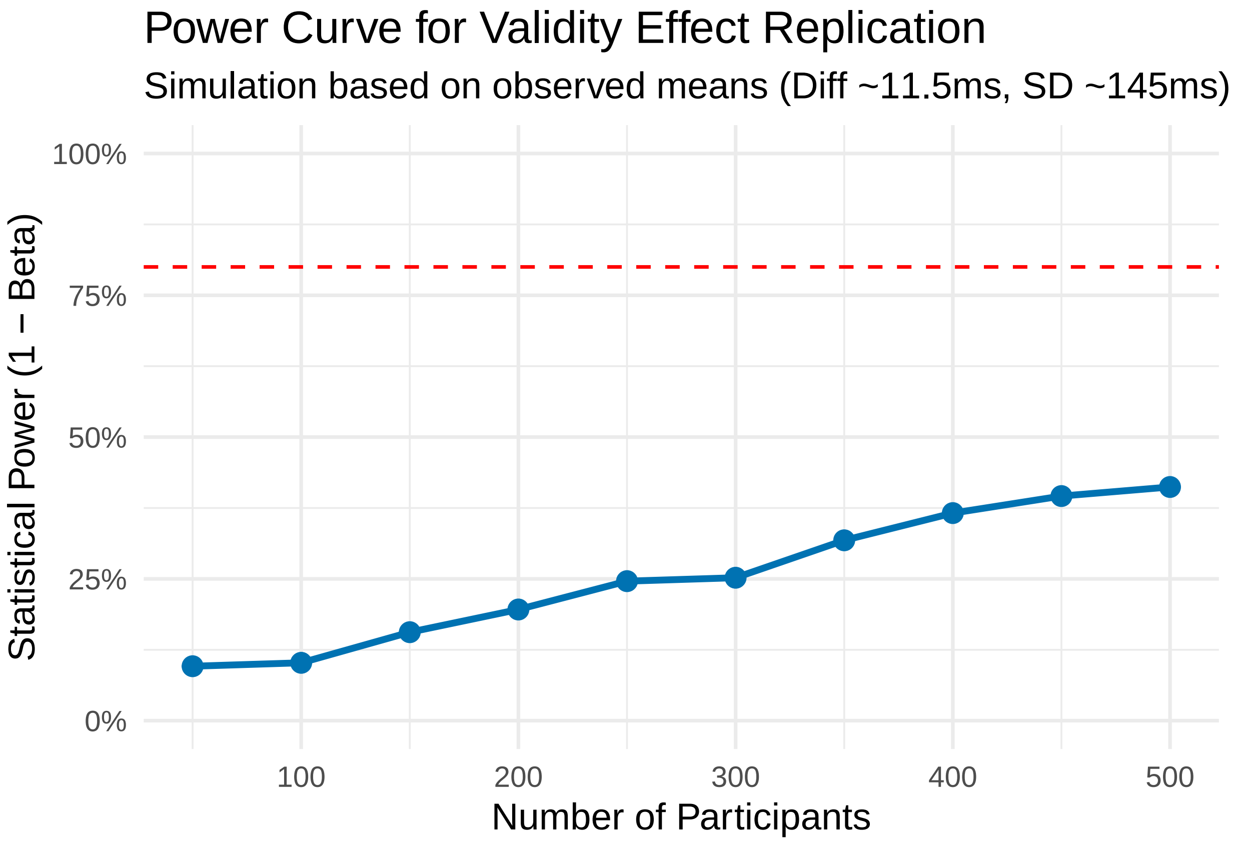
Next, to statistically assess whether accuracy differed by cue type, cue validity, or their interaction, I have a binomial logistic regression. Logistic models are appropriate for binary outcomes here. The model included cueType, valid, and their interaction as predictors.

```
##
## Call:
## glm(formula = correct ~ cueType * valid, family = binomial, data = dat)
##
```

```
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                3.04397    0.10974  27.739   <2e-16 ***
## cueTypeface                0.12974    0.15991   0.811    0.417
## validinvalid              -0.04495    0.15355  -0.293    0.770
## cueTypeface:validinvalid  0.09775    0.22661   0.431    0.666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2708.6  on 7658  degrees of freedom
## Residual deviance: 2705.9  on 7655  degrees of freedom
##   (3836 observations deleted due to missingness)
## AIC: 2713.9
##
## Number of Fisher Scoring iterations: 6
```

Consistent with the descriptive plot, the logistic regression revealed **no significant main effects of cue type or validity, and no interaction between them**. All predictors had p-values well above .40, indicating that neither cue type nor cue validity reliably influenced accuracy. This is very normal here; This confirms that accuracy remained stable across conditions and suggests that participants were equally capable of performing the task regardless of cue direction or other things.

simulation-based power analysis To evaluate how many participants would be required to reliably detect the validity effect, I conducted a simulation-based power analysis. I use empirical mean difference between valid and invalid trials, the observed variability (approximate 145 ms), and a moderate within-subject correlation, I simulated datasets of different sample sizes and tested each using a paired t-test.500 simulated experiments were run.



The results of the simulated power analysis shows that detecting the validity effect observed in the present dataset would require a very large sample size. Power would probably exceed 50% until approximately 500 participants, suggesting that it would be difficult to detect in typical laboratory samples.

Discussion

The re-analysis of Experiment 1 data bere reproduced the core validity effects typically observed in attentional cueing paradigms; that is people would respond faster when cue is validly pointing to the target. Although the effects were small and statistically weak, I believe people would observe a better effect size statistics in the other experiment.

Reaction times were slightly faster on valid than invalid trials, but the difference was modest and did not significantly interact with cue type.

In contrast to the original paper, which had a much richer set of experiments and larger sample sizes, Experiment 1 alone provided limited evidence for strong attentional advantages, likely because the design was simple and individual differences were not modeled (People are doing extremely great on all trials).Therefore, from this perspective, **I succesffuly reproduced the attentional effect, but in weak form.** Overall, these results suggest that while the validity effect is present, it is not robust enough in this dataset to support broader theoretical conclusions without additional data.

Then I did analysis to examine the statistical power required to detect the observed effects. The simulation-based power analysis used the empirically estimated effect size from the current dataset and repeatedly simulated experiments of varying sample sizes. These simulations showed that the effect of cue validity was very small relative to the trial-to-trial variability in reaction times. As a result, the model estimated that extremely large sample sizes (at least 500 participants)would be needed to achieve conventional levels of statistical power. This finding is consistent with the above analyses and emphasize that the present data are underpowered for detecting subtle attentional differences, especially interactions.

These power results also shows that the experimental design might be improved to achieve stronger or more reliable effects. Increasing the number of participants is the most straightforward solution here (for instance, if the experiment 1 has 500 participants, it would lead to very promising results). I think there are potential other solutions that can be involved other than increasing the number of participants. For example, we could use more trials per condition, this would reduce within-participant variability. In summary, the results of the above analysis and the power simulations illustrate that while attentional cueing effects are theoretically robust, detecting them reliably would need adequate sample size.

References

Schmitz, Strauss, Reinel, and Einhäuser (2024)

Schmitz, I., Strauss, H., Reinel, L., & Einhäuser, W. (2024). Attentional cueing: Gaze is harder to override than arrows. *PLOS ONE*, *19*(3), e0301136. https://doi.org/10.1371/journal.pone.0301136