

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

3 Yujingai (Emmy) Shi¹

4 ¹ Princeton University

5 Author Note

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

8

Abstract

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

17 *Keywords:* statistics, eyetracking, attention tests, psychophysics

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

20 **Introduction**

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

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

38 **Method**

39 Below I am listing the Experiment 1 set up and procedure, and the data analysis
40 pipeline I used to examine the attentional effects and the power tests. ## Procedure
41 Experiment 1 used a spatial cueing paradigm; each trial began with either an eye-gaze cue
42 (a face whose eyes shifted left or right) or an arrow cue pointing in one direction. After a

43 brief cue-target interval, a target letter appeared on either the left or right side of the
 44 display, and participants responded by pressing a key to record. On valid trials the cue
 45 correctly indicated the target location; on invalid trials it pointed in the opposite direction.
 46 Reaction time and accuracy were recorded on every trial. All data for Experiment 1 were
 47 publicly available through the OSF repository associated with the original article. ##
 48 Data analysis In this project, I re-analyzed Experiment 1 using the statistical tools learned
 49 in PSY503. The primary analyses focused on computing the classic validity effect (valid
 50 vs. invalid trials) and testing whether this effect differed between gaze cues and arrow cues
 51 through ANOVA and Regression tests. I also conducted supplementary analyses, including
 52 visualization of reaction time and accuracy performance, and a simulation-based power
 53 analysis to assess the sample size required to detect the observed effect sizes.

54 **Results**

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

```

57 ## Warning: package 'ggplot2' was built under R version 4.5.2

58 ##   VP block trial cueType      cue tgt tgtLoc resp     RT valid train valid2
59 ## 1   1     1     1   face   right    2  right     A 1218  TRUE train  TRUE
60 ## 2   1     1     2   face   left     1  right     B  656 FALSE train FALSE
61 ## 3   1     1     3   face   left     2  right     A  445 FALSE train FALSE
62 ## 4   1     1     4   face neutral   2  right     A  371 FALSE train    NA
63 ## 5   1     1     5   face  right    1  left      B  511 FALSE train FALSE
64 ## 6   1     1     6   face  right    2  left      A  527 FALSE train FALSE
65 ##   instruction tgtResp correct      cueDir      tgtDir
66 ## 1             A          2  TRUE horizontal horizontal
67 ## 2             A          1  TRUE horizontal horizontal
  
```

```
68 ## 3           A     2   TRUE horizontal horizontal
69 ## 4           A     2   TRUE      neutral horizontal
70 ## 5           A     1   TRUE horizontal horizontal
71 ## 6           A     2   TRUE horizontal horizontal
```

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

75 And we can see that we have 3830 valid trials where cue predicted the target and
76 3829 invalid trials where cue misled the target, so this means the project has balanced
77 trials, which is good for further analysis.

```

94 ## $ correct      <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE~
95 ## $ cueDir       <chr> "horizontal", "neutral", "neutral", "horizontal", "neutral~
96 ## $ tgtDir       <chr> "horizontal", "horizontal", "horizontal", "horizontal", "h~
97 ##
98 ##   valid invalid
99 ##   3830     3829

100 ##   Min. 1st Qu. Median    Mean 3rd Qu.    Max.
101 ##   179.0  370.0  432.0  460.5  510.0  1983.0

```

102 Descriptive Statistics: to compute descriptive statistics by cuetype and validity
103 Before the main analyses, we could compute **summary statistics** for each condition (cue
104 type \times validity). This helps check whether the expected validity effect appears in both cue
105 types. For each condition,I computed the mean RT(reaction time), standard deviation,
106 sample size, and the standard error of the mean (SEM).

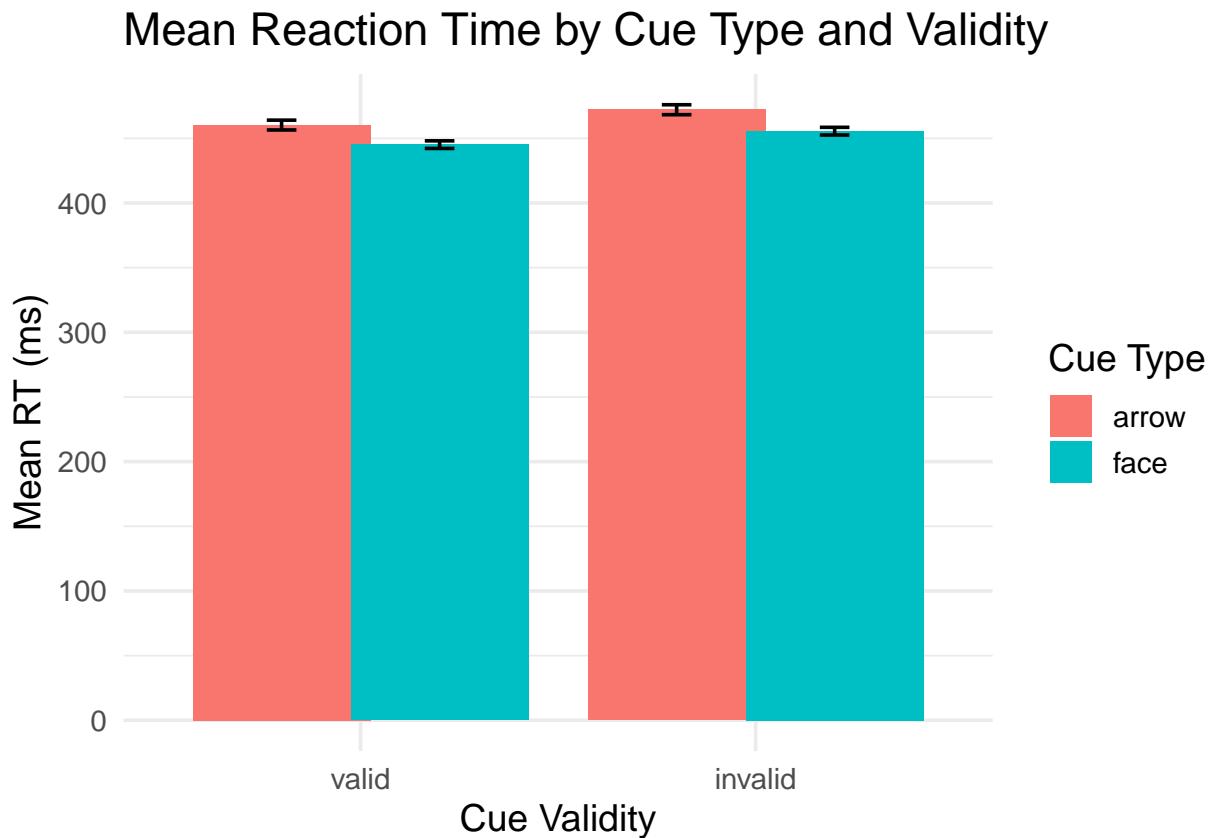
```

107 ## # A tibble: 4 x 6
108 ##   cueType valid   mean_RT  sd_RT      n  se_RT
109 ##   <fct>    <fct>    <dbl> <dbl> <int> <dbl>
110 ## 1 arrow    valid    460.  166.  1913  3.79
111 ## 2 arrow    invalid  472.  168.  1917  3.85
112 ## 3 face    valid    445.  129.  1917  2.94
113 ## 4 face    invalid  456.  130.  1912  2.97

```

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

¹¹⁸ Below is the visualization for the descriptive data



¹¹⁹

¹²⁰ This plot shows mean reaction times for valid versus invalid cues, for arrow cues and
¹²¹ face(gaze) cues. Face cues produce overall faster reaction times than arrow cues, and valid
¹²² trials are faster than invalid trials.

¹²³ ****2 *2 ANOVA on Reaction Time**** Now I want to specifically know that 1) Are
¹²⁴ reaction times faster on valid versus invalid trials?, and 2) Does the size of the validity
¹²⁵ effect differ for arrow versus face cues? So I have a 2x2 ANOVA. I fit a two-way ANOVA
¹²⁶ predicting RT from cue type (arrow vs face), cue validity (valid vs invalid), and their
¹²⁷ interaction. Only trials with clear validity labels (valid/invalid) are included here.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
¹²⁹ ## cueType	1	483078	483078	21.647	3.33e-06 ***
¹³⁰ ## valid	1	237833	237833	10.658	0.0011 **
¹³¹ ## cueType:valid	1	1039	1039	0.047	0.8292

```

132 ## Residuals      7655 170828875   22316
133 ## ---
134 ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
135 ## 3836 observations deleted due to missingness

```

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

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

```

145 ## # Effect Size for ANOVA (Type I)
146 ##
147 ## Parameter | Eta2 (partial) | 95% CI
148 ## -----
149 ## cueType | 2.82e-03 | [0.00, 1.00]
150 ## valid | 1.39e-03 | [0.00, 1.00]
151 ## cueType:valid | 6.08e-06 | [0.00, 1.00]
152 ##
153 ## - One-sided CIs: upper bound fixed at [1.00].

```

154 Although the partial eta-squared values are small ($< .01$), this pattern is typical for
 155 reaction time data, where variability is influenced by many cognitive and motor processes.I
 156 think that small effect sizes do not indicate a problem with the analysis; rather, they reflect

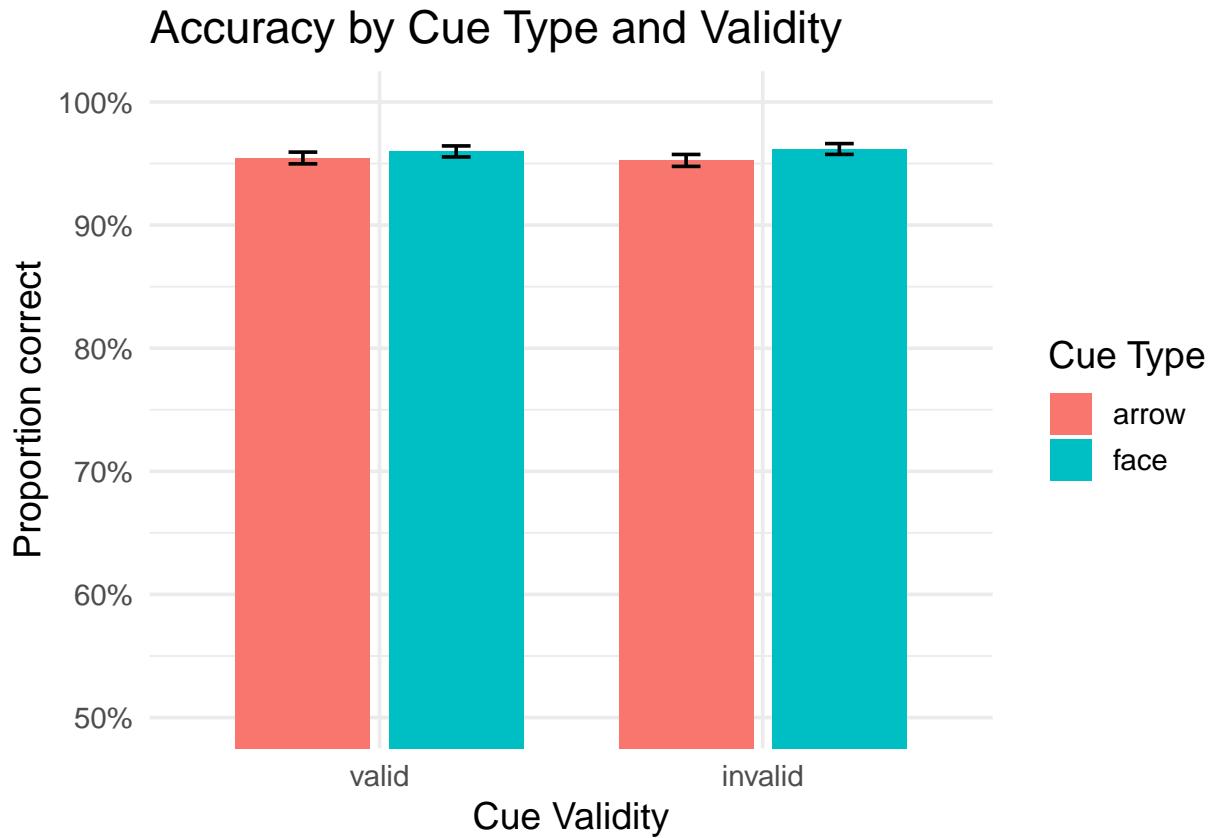
157 that the experimental manipulations account for a modest portion of RT variance. Finally,
158 this is consistent with what is commonly observed in attention research.

159 Now I have calculated the accuracy for the different cue types and validity status.

160 Next, I am going to perform analysis on accuracy.

```
161 ## # A tibble: 4 x 5
162 ##   cueType valid  mean_acc     n   se_acc
163 ##   <fct>    <fct>    <dbl> <int>   <dbl>
164 ## 1 arrow    valid     0.955  1913  0.00476
165 ## 2 arrow    invalid   0.953  1917  0.00486
166 ## 3 face    valid     0.960  1917  0.00448
167 ## 4 face    invalid   0.962  1912  0.00438
```

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



173

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

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

182 ##

183 ## Call:

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

185 ##

```

186 ## Coefficients:
187 ## Estimate Std. Error z value Pr(>|z|)
188 ## (Intercept) 3.04397 0.10974 27.739 <2e-16 ***
189 ## cueTypeface 0.12974 0.15991 0.811 0.417
190 ## validinvalid -0.04495 0.15355 -0.293 0.770
191 ## cueTypeface:validinvalid 0.09775 0.22661 0.431 0.666
192 ## ---
193 ## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
194 ##
195 ## (Dispersion parameter for binomial family taken to be 1)
196 ##
197 ## Null deviance: 2708.6 on 7658 degrees of freedom
198 ## Residual deviance: 2705.9 on 7655 degrees of freedom
199 ## (3836 observations deleted due to missingness)
200 ## AIC: 2713.9
201 ##
202 ## Number of Fisher Scoring iterations: 6

```

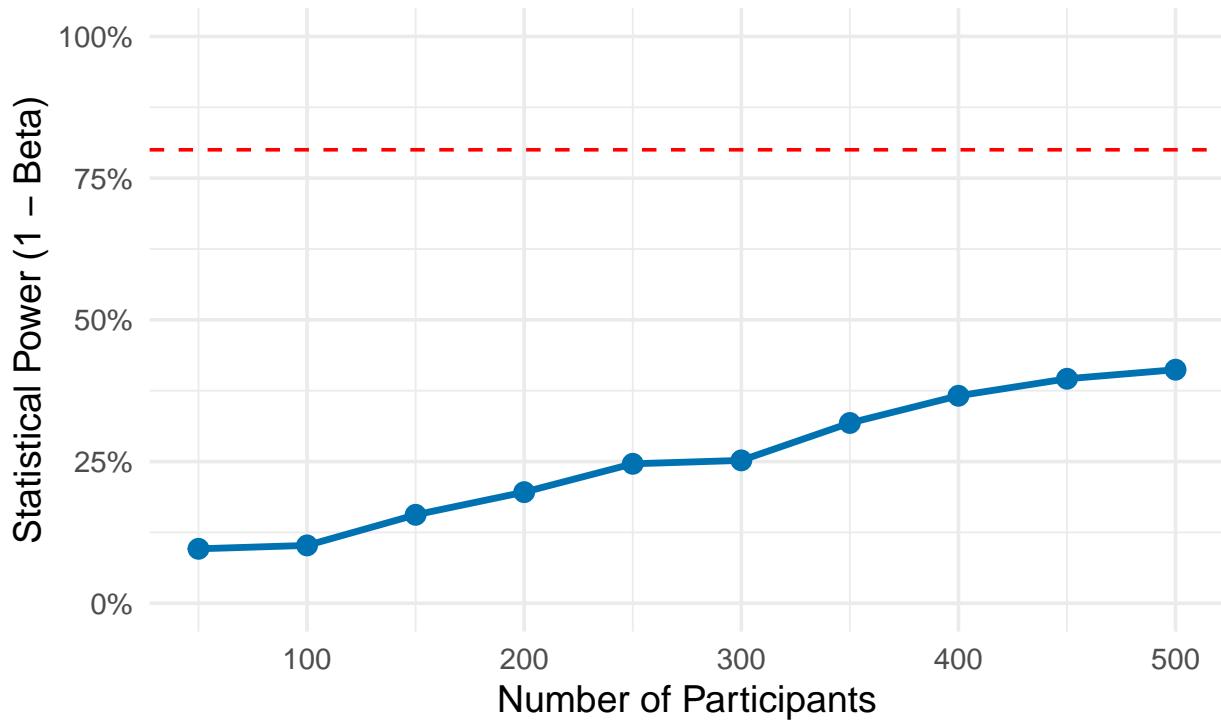
203 Consistent with the descriptive plot, the logistic regression revealed **no significant**
 204 **main effects of cue type or validity, and no interaction between them.** All
 205 predictors had p-values well above .40, indicating that neither cue type nor cue validity
 206 reliably influenced accuracy.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

209 **simulation-based power analysis** To evaluate how many participants would be
 210 required to reliably detect the validity effect, I conducted a simulation-based power
 211 analysis. I use empirical mean difference between valid and invalid trials, the observed

²¹² variability (approximate 145 ms), and a moderate within-subject correlation, I simulated
²¹³ datasets of different sample sizes and tested each using a paired t-test.⁵⁰⁰ simulated
²¹⁴ experiments were run.

Power Curve for Validity Effect Replication

Simulation based on observed means (Diff ~11.5ms, SD ~145ms)



²¹⁵

²¹⁶ The results of the simulated power analysis shows that detecting the validity effect
²¹⁷ observed in the present dataset would require a very large sample size. Power would
²¹⁸ probably exceed 50% until approximately 500 participants, suggesting that it would be
²¹⁹ difficult to detect in typical laboratory samples.

²²⁰

Discussion

²²¹ The re-analysis of Experiment 1 data bere reproduced the core validity effects
²²² typically observed in attentional cueing paradigms; that is people would respond faster
²²³ when cue is validly pointing to the target. Although the effects were small and statistically
²²⁴ weak, I believe people would observe a better effect size statistics in the other experiment.

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

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

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

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

251

References

252 Schmitz, Strauss, Reinel, and Einhäuser (2024)

253 Schmitz, I., Strauss, H., Reinel, L., & Einhäuser, W. (2024). Attentional cueing: Gaze is
254 harder to override than arrows. *PLOS ONE*, 19(3), e0301136.255 <https://doi.org/10.1371/journal.pone.0301136>