Supplementary analyses

This document details additional analyses that were conducted to 1) assess the size and stability of the confidence interval of our analytic over time, 2) compare the difficulty of conducive and non-conducive problems when they are both solved through a left-to-right procedure, and 3) assess whether removing the false alarms altered the results of the studies. Each analysis is reported below.

# 1. Confidence interval over time

We reanalysed our data in two ways to assess the size and the stability of the confidence interval described in our manuscript (referred to here as the ‘original metric’). In the first analysis we implemented a new metric where the confidence interval around the non-conducive mean was derived from three trials (to match the rolling median of conducive trials). We compared the data that the new metric generated with that derived from our original metric. In the second analysis we compared the size of the confidence interval at the mid-point of the trial sequence to the size of the confidence interval at the end of the trial sequence. The full details of the analysis are reported below, and we draw four conclusions from it:

a) Most importantly, the use of a different metric does not alter the outcome of any of our pre-registered analyses comparing the attention conditions. Therefore, this specific detail of the method has not influenced our substantive conclusions.

b) The first analysis reported below shows that the new metric reduces the number of false alarms and increases the number of misses. The new metric reduced the number of false alarms approximately by a factor of 5, but it increased the number of misses by a factor of 6, suggesting that that it does not improve our analytic overall.

c) The second analysis showed that the size of the confidence interval reduced over time, but the value of the lower-bound of the confidence interval was similar at the two time-points. This is likely to be because people gradually improve in solving non-conducive problems over time, so any reduction in the width of the CI is offset by the lower value of the mean, making the absolute value of the lower-bound relatively stable.

d) There are numerous ways that a confidence interval could be calculated (e.g. from a mean of 4 trials, 5 trials, 10 trials), and we judge that each will have a trade-off between sensitivity (the ability to detect people who do identify) and specificity (the ability to correctly reject people who do not identify). In our studies we valued sensitivity in that we did not want to miss an individual who identified the shortcut. We therefore believe that the original way that we defined the metric was most appropriate for our studies. We acknowledge however that other studies wishing to use our metric could adjust the criteria of it to their needs, for example, by having a confidence interval generated from a mean of three trials, or by applying a retrospective confidence interval to the trials after all have been completed. For our studies, the metric was generated in live-time and we judge that by including all of the non-conducive trials that had been presented at each time-point was the most accurate way for inferring identification.

## Analysis 1.1: A new metric based on a rolling confidence interval of 3 non-conducive problems

The aim of our first analysis was to see if applying a confidence interval using a new metric altered our results, and if so, by how much.

On each trial, the mean of the three most recent non-conducive trials and the 99.9% confidence interval of the mean were calculated. Our pre-registered dependent variables were then recalculated, and our analyses repeated. The results are reported below, and are compared to the findings derived from the original metric in our manuscript.

### Study 1

#### False alarms and misses

There were 3 false alarms and 5 misses using the new metric, compared to 15 false alarms and one miss using the original metric.

#### Number of identifiers

Table 1 displays the number of identifiers and non-identifiers using the new metric, compared to the old metric.

Table 1: Number of identifiers and non-identifiers of a shortcut using the new metric and the original metric in each of the three conditions of Study 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Condition |  |  |  |
|  | Left-prime | Control | Right-prime | Total |
|  | Original metric | New metric | Original metric | New metric | Original metric | New metric | Original metric | New metric |
| Identifier | 19 | 17 | 17 | 17 | 25 | 23 | 61 | 57 |
| Non-identifier | 17 | 19 | 19 | 19 | 11 | 13 | 47 | 53 |
| Total | 36 | 36 | 36 | 36 | 36 | 36 | 108 | 108 |

The relative frequency of identifiers to non-identifiers was not significantly different across the three conditions, *x2*(2) = 2.67, *p* = 0.262, Cramer’s V = 0.16.

#### Trial numbers of the identification point

The new metric changed the trial number of identification for 29 of the 57 identifiers. For 22 of these individuals, the new metric made their trial number later than that calculated by the original metric. The mean difference in the trial number of identification between the new and original metric was ±3.10 (SD = 2.43, range = 1 – 9).

Table 2 displays descriptive statistics of the trial number of identification using the original metric and the new metric.

Table 2: Mean (SD) trial number of identification using the new metric and original metric in each of the three conditions of Study 1.

|  |  |
| --- | --- |
|  | Condition |
|  | Left-prime | Control | Right-prime |
|  | Original metric | New metric | Original metric | New metric | Original metric | New metric |
| Conducive trial number | 8.00 (3.73) | 8.35 (4.49) | 6.71 (3.60) | 8.77 (4.52) | 6.44 (5.36) | 6.17 (4.26) |

The data were not normally distributed. A Kruskal-Wallis test with three levels (left-prime, right-prime, control) found no significant difference between the conditions in the conducive trial number of identification, *x2*(2) = 4.54, *p* = 0.10, or the total trial number of identification, *x2*(2) = 4.54, *p* = 0.10.

#### Response time across all trials

For those who were classed as identifiers, median RT of the correctly solved trials were calculated. A 3\*2 mixed ANOVA with condition (left-prime, right-prime, control) and problem type (conducive, non-conducive) as between and within-subject factors found a main effect of problem type, *F*(1, 54) = 175.39, *p*<0.001, ηp2 = 0.76, where conducive problems were solved quicker than non-conducive problems. There was no main effect of condition, *F*(2, 54) = 1.03, *p =* 0.365, ηp2 = 0.04, and no significant interaction between problem type and condition, *F*(2, 54) = 0.53, *p =* 0.592, ηp2 = 0.02.

#### Response time for the trials after the identification point

For those who were classed as identifiers, median RT of the correctly solved trials after the IP were calculated. The data were analysed for a) statistical equivalence and b) statistical difference, with adjusted alpha levels for the number of comparisons (0.05/6 = 0.008). Table 3 displays the result. None of the comparisons were statistically equivalent (*p* > 0.008 for all comparisons) and none were significantly different.

Table 3: Outcome of the equivalence tests and tests of statistical difference for median RT (s) after the IP. The data are from the 17, 17 and 23 identifiers in the control, left-prime and right-prime conditions respectively[[1]](#footnote-1).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Problem type | Comparison | Mean (SD) | Equivalence test result | Statistical difference result | Interpretation (with alpha adjusted to *p*< 0.008) |
| Conducive | Control and left-prime | 5.30 (2.41) and 3.72 (1.72) | *t*(32) = 0.13, *p* = 0.448 | *t*(32) = 2.20, *p* = 0.035 | Not statistically equivalent or statistically different |
| Control and right-prime  | 5.30 (2.41) and 5.79 (6.04) | *t(*38) = 2.19, *p* = 0.018 | *t*(38) = 0.32, *p* = 0.754 | Not statistically equivalent or statistically different |
| Left-prime and right-prime  | 3.72 (1.72) and 5.79 (6.04) | *t*(22.97) = 1.98, *p* = 0.030 | *t*(38) = 1.37, *p* = 0.179 | Not statistically equivalent or statistically different |
| Non-conducive | Control and left-prime  | 13.92 (6.99) and 9.86 (4.28) | *t*(32) = 0.29, *p* = 0.387 | *t*(32) = 2.04, *p* = 0.049 | Not statistically equivalent or statistically different |
| Control and right-prime | 13.92 (6.99) and 12.21 (5.91) | *t*(38) = 1.66, *p* = 0.052 | *t*(38) = 0.84, *p* = 0.405 | Not statistically equivalent or statistically different |
| Left-prime and right-prime  | 9.86 (4.28) and 12.21 (5.91) | *t*(38) = 1.11, *p* = 0.137 | *t*(38) = 1.39, *p* = 0.174 | Not statistically equivalent or statistically different |

#### Accuracy across all trials

For those who were classed as identifiers, the mean percent of correctly solved problems across all trials were calculated. A 3\*2 mixed ANOVA with condition (left-prime, right-prime, control) and problem type (conducive, non-conducive) as between and within-subject factors found a main effect of problem type, *F*(1, 54) = 33.13, *p*<0.001, ηp2 = 0.38, where conducive problems were solved more accurately than non-conducive problems. There was no main effect of condition, *F*(2, 54) = 0.41, *p =* 0.667, ηp2 = 0.01, and no interaction between problem type and condition, *F*(2, 54) = 0.19, *p =* 0.826, ηp2 < 0.01.

#### Accuracy for the trials after the identification point

For those who were classed as identifiers, accuracy of the trials after the IP were calculated. The data were analysed for a) statistical equivalence and b) statistical difference, with adjusted alpha levels for the number of comparisons (0.05/6 = 0.008). Table 4 displays the result. None of the comparisons were statistically equivalent (*p* > 0.008 for all comparisons) and none were significantly different.

Table 4: Outcome of the equivalence tests and tests of statistical difference for solution accuracy (%) after the IP. The data are from the 17, 17 and 23 identifiers in the control, left-prime and right-prime conditions respectively[[2]](#footnote-2).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Problem type | Comparison | Mean (SD) | Equivalence test result | Statistical difference result | Interpretation (with alpha adjusted to *p*< 0.008) |
| Conducive | Control and left-prime | 92.35 (13.30) and 91.86 (10.33) | *t*(32) = 2.21, *p* = 0.017 | *t*(32) = 0.12, *p* = 0.904 | Not statistically equivalent or statistically different |
| Control and right-prime | 92.35 (13.30) and 97.18 (4.62) | *t*(38) = 0.88, *p* = 0.192 | *t*(38) = 1.61, *p* = 0.114 | Not statistically equivalent or statistically different |
| Left-prime and right-prime | 91.86 (10.33) and 97.18 (4.62) | *t*(20.75) = 0.40, *p* = 0.346 | *t*(38) = 2.20, *p* = 0.034 | Not statistically equivalent or statistically different |
| Non-conducive | Control and left-prime | 83.86 (19.72) and 76.98 (25.61) | *t*(32) = 1.45, *p* = 0.078 | *t*(32) = 0.88, *p* = 0.387 | Not statistically equivalent or statistically different |
| Control and right-prime | 83.86 (19.72) and 82.84 (13.46) | *t*(38) = 2.31, *p* = 0.013 | *t*(38) = 0.19, *p* = 0.847 | Not statistically equivalent or statistically different |
| Left-prime and right-prime | 76.98 (25.61) and 82.84 (13.46) | *t*(38) = 1.56, *p* = 0.063 | *t*(38) = 0.94, *p* = 0.354 | Not statistically equivalent or statistically different |

### Study 2

#### False alarms and misses

There were 4 false alarms and 6 misses with the new metric, compared to 13 false alarms and one miss with the original metric.

#### Number of identifiers

Using the new metric, there were 50 identifiers and 58 non-identifiers. Table 5 displays the number of identifiers and non-identifiers using the new metric, compared to the old metric.

Table 5: Number of identifiers and non-identifiers of a shortcut using the new metric and the original metric in each of the three conditions of Study 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Condition |  |  |  |
|  | Left-prime | Control | Right-prime | Total |
|  | Original metric | New metric | Original metric | New metric | Original metric | New metric | Original metric | New metric |
| Identifier | 18 | 16 | 17 | 14 | 20 | 20 | 55 | 50 |
| Non-identifier | 18 | 20 | 19 | 22 | 16 | 16 | 53 | 58 |
| Total | 36 | 36 | 36 | 36 | 36 | 36 | 108 | 108 |

The relative frequency of identifiers to non-identifiers was not significantly different across the three conditions, *x2*(2) = 2.09, *p* = 0.352, Cramer’s V = 0.14.

#### Number of early identifiers

Table 6 displays the number of early identifiers and non-early identifiers using the new metric, compared to the original metric.

Table 6:Number of early identifiers and non-early identifiers of a shortcut using the new metric and the original metric in each of the three conditions of Study 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Condition |  |  |  |
|  | Left-prime | Control | Right-prime | Total |
|  | Original metric | New metric | Original metric | New metric | Original metric | New metric | Original metric | New metric |
| Early identifier | 3 | 2 | 5 | 4 | 5 | 4 | 13 | 10 |
| Non early-identifier | 33 | 34 | 31 | 32 | 31 | 32 | 95 | 98 |
| Total | 36 | 36 | 36 | 36 | 36 | 36 | 108 | 108 |

The relative frequency of early identifiers to non-early identifiers was not significantly different across the three conditions, *x2*(2) = 0.882, *p* = 0.644, Cramer’s V = 0.09.

#### Percent of conducive trials after the identification point

The new metric changed the trial number of identification for 31 of the 50 identifiers. For 23 of these individuals, the new metric made their trial number later than their trial number according to the original metric. The mean difference in the trial number of identification between the new and original metric was ±3.90 (SD = 3.39, range = 1 – 11).

As detailed in our manuscript, the trial number of identification was converted into the percent of trials remaining after the identification point. Table 7 compares the percent of conducive trials after identification point using the original and new metric.

Table 7: Mean (SD) trial number of identification using the new metric and original metric in each of the three conditions of Study 1.

|  |  |
| --- | --- |
|  | Condition |
|  | Left-prime | Control | Right-prime |
|  | Original metric | New metric | Original metric | New metric | Original metric | New metric |
| Percent of trials after the IP | 40.53 (42.58) | 34.09 (40.34) | 36.24 (41.92) | 32.95 (42.69) | 44.07 (41.95) | 38.76 (39.81) |

The data were not normally distributed, and a Kruskal-Wallis test using all participants’ data (identifiers and non-identifiers) found no significant difference between the conditions, *X*2(2) = 0.68, *p* = 0.713.

#### Reaction time and accuracy for all of the trials

This analysis is unaltered by the criteria of the metric. Therefore, the outcome is identical to what is reported in our manuscript.

## Analysis 1.2: Comparison of the 99.9% confidence interval at the mid-way and end of the trials

The aim of our second analysis was to assess the stability of the confidence interval generated by the original metric in our manuscript.

The plots below depict the conducive median (black), the non-conducive mean (red) and the 99.9% confidence interval (blue) of 3 identifiers (left panels) and 3 non-identifiers (right panels).

 ****

****

****

The graphs illustrate that:

* The lower-bound of the confidence interval is relatively stable by the half-way point (non-conducive trial number 12).
* Although the width of the confidence interval reduces with more trials, the absolute value of the lower-bound is relatively stable because the non-conducive mean often gradually declines.
* Identification of the shortcut is accompanied by a change in RT that is much larger than any gentle change in the confidence interval of the non-conducive problems.

We re-analysed the data by calculating a) the size of the confidence interval and b) the absolute value of the lower-bound of the confidence interval at the half-way point (non-conducive trial number 12, trial number 23) and at the end of the experiment (non-conducive trial number 23, trial number 45). Table 8 displays the result.

Table 8: Size of the confidence interval and value of the lower-bound at the half-way point and end of the trial sequence.

|  |  |  |
| --- | --- | --- |
|  | Confidence interval | Lower-bound of non-conducive mean |
|  | Mid-way point | End of trial sequence | Mid-way point | End of trial sequence |
| Study 1 | 5.05 (3.33) | 3.74 (2.31) | 10.84 (5.08) | 11.07 (5.17) |
| Study 2 | 6.33 (4.74) | 4.41 (2.89) | 12.16 (5.10) | 12.48 (4.81) |

For Study 1, there was a significant difference in the size of the confidence interval at the two time-points, *W* (108) = 5553.00, *p* < 0.001. There was no significant difference in the value of the lower-bound of the confidence interval, *W* (108) = 2564.00, *p* = 0.256.

For Study 2, there was a significant difference in the size of the non-conducive confidence interval at the two time-points, *W* (108) = 5688.00, p < 0.001. There was also a significant difference in the value of the lower-bound of the confidence interval, *t*(107) = 2.10, *p* = 0.038.

# 2. Comparing the difficulty of conducive and non-conducive problems when solved through a left-to-right procedure:

To the best of our knowledge, there is no known factor that would make conducive problems systematically easier than non-conducive problems if both problem types are solved through a left-to-right procedure. The problems were carefully created, and match on the size of the result of the interim addition (a + b), the size of the subtraction (c) and the answer. They also match on the number and size of carry operations in the addition. We judge that they are as closely matched as possible without them being the same problem.

To evidence that this is the case, we re-analysed our data. We compared the accuracy and median correct response time for solving the two problem types by the self-reported non-identifiers (i.e. participants who did not use the associativity shortcut strategy at any point in the experiment). If the problems are similar in difficulty when they are not solved through a shortcut, there should be no significant difference between them. Table 9 details the result. Indeed, in Study 1, Study 2 and the pilot study, there was no difference between the problem types on either outcome measure. Moreover, equivalence tests demonstrate that the two problem types were statistically equivalent. The analyses are reported below.

## Study 1

There was no significant difference in the percent of correctly answered conducive and non-conducive problems, *t*(45) = 1.361, *p* = 0.180 and no significant difference in the median RT of correctly answered conducive and non-conducive problems, *t*(45) = 0.238, *p* = 0.813. Moreover, the accuracy scores were statistically equivalent using a test of one sided significance, *t*(45) = 2.03, *p* = 0.024. The RT scores were also statistically equivalent, *t*(45) = 3.16, *p* = 0.001.

## Study 2

There was no significant difference in the percent of correctly answered conducive and non-conducive problems, *t*(51) = 0.394, *p* = 0.695 and no significant difference in the median RT of correctly answered conducive and non-conducive problems, *t*(51) = 0.469, *p* = 0.641. Moreover, the accuracy scores were statistically equivalent using a test of one sided significance, *t*(51) = 3.21, *p* = 0.001. The RT scores were also statistically equivalent, *t*(51) = 3.42, *p* = 0.001.

## Pilot study

There was no significant difference in the percent of correctly answered conducive and non-conducive problems, *t*(7) = 0.53, *p* = 0.612, and no significant difference in the median RT of correctly answered conducive and non-conducive problems, *t*(7) = 0.97, *p* = 0.364. The accuracy scores were not statistically equivalent using a test of one sided significance, *t*(7) = 0.89, *p* = 0.203. The RT scores were not statistically equivalent, *t*(7) = 1.41, *p* = 0.101.

Table 9: Descriptive statistics and outcomes of statistical significance and equivalence tests on accuracy and RT data for self-reported non-identifiers of the shortcut.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Accuracy (%) | Outcome of accuracy analyses | Median RT (s) | Outcome of RT analyses |
| Study 1 | ConduciveNon-conducive | 80.91 (14.48)83.08 (12.54) | Not significantly different Statistically equivalent | 15.42 (6.82)15.35 (7.05) | Not significantly different Statistically equivalent |
| Study 2 | ConduciveNon-conducive | 83.19 (13.48)83.70 (11.83) | Not significantly different Statistically equivalent | 16.68 (6.15)16.57 (6.08) | Not significantly different Statistically equivalent |
| Pilot study | ConduciveNon-conducive | 76.47 (31.71)74.60 (29.65) | Not significantly different Not statistically equivalent | 15.77 (3.64)16.68 (4.36) | Not significantly different Not statistically equivalent |

We performed the same analyses described above on the accuracy and RT data for self-reported identifiers. In Study 1, there was a significant difference in the percent of correctly answered conducive and non-conducive problems, *t*(61) = 5.94, *p* <0.001, and a significant difference in the median RT of correctly answered conducive and non-conducive problems, *t*(61) = 12.48, *p*< 0.001. In Study 2, there was a significant difference in the percent of correctly answered conducive and non-conducive problems, *t*(55) = 6.86, *p* <0.001, and a significant difference in the median RT of correctly answered conducive and non-conducive problems, *t*(55) = 12.83, p < 0.001.

# 3. Analysis of false alarms

We reanalysed our data with the false alarms classed as identifiers. The pattern of results stayed the same as the results in our manuscript. Full details are reported below.

## Study 1

#### Number of identifiers

There was no significant difference in the number of identifiers across the three conditions, *x2*(2) = 2.75, *p* = 0.252, Cramer’s V = 0.16.

#### Trial numbers of the identification point

The data were not normally distributed. A Kruskal-Wallis test with three levels (left-prime, right-prime, control) found no significant difference between the conditions in the conducive trial number of identification, *X2*(2) = 2.72, *p* = 0.257, or the total trial number of identification, *X2*(2) = 2.72, *p* = 0.257.

#### Response time across all trials

A 3\*2 mixed ANOVA with condition (left-prime, right-prime, control) and problem type (conducive, non-conducive) as between and within-subject factors found a main effect of problem type, *F*(1, 73) = 112.56, *p*<0.001, ηp2 = 0.61, where conducive problems were solved quicker than non-conducive problems. There was no main effect of condition, *F*(2, 73) = 1.80, *p =* 0.172, ηp2 = 0.05, and no significant interaction between problem type and condition, *F*(2, 73) = 0.19, *p =* 0.832, ηp2 = 0.01.

#### Response time for the trials after the identification point

For those who were classed as identifiers, median RT of the correctly solved trials after the IP were calculated. The data were analysed for a) statistical equivalence and b) statistical difference, with adjusted alpha levels for the number of comparisons (0.05/6 = 0.008). None of the comparisons were statistically equivalent (*p* > 0.008 for all comparisons) and none were significantly different.

#### Accuracy across all trials

For those who were classed as identifiers, the mean percent of correctly solved problems across all trials were calculated. A 3\*2 mixed ANOVA with condition (left-prime, right-prime, control) and problem type (conducive, non-conducive) as between and within-subject factors found a main effect of problem type, *F*(1, 73) = 22.86, *p*<0.001, ηp2 = 0.24, where conducive problems were solved more accurately than non-conducive problems. There was no main effect of condition, *F*(2, 73) = 0.41, *p =* 0.668, ηp2 = 0.01, and no interaction between problem type and condition, *F*(2, 73) = 0.46, *p =* 0.632, ηp2 < 0.01.

#### Accuracy for the trials after the identification point

For those who were classed as identifiers, accuracy of the trials after the IP were calculated. The data were analysed for a) statistical equivalence and b) statistical difference, with adjusted alpha levels for the number of comparisons (0.05/6 = 0.008). One comparison was statistically equivalent, the control and right-prime condition in their accuracy scores on the non-conducive problems after their IP, *t*(51) = 2.89, *p* = 0.003. None of the other comparisons were statistically equivalent. None of the comparisons were significantly different.

## Study 2

#### Number of identifiers

There was no significant difference in the number of identifiers across the three conditions, *x2*(2) = 0.56, *p* = 0.757, Cramer’s V = 0.07.

#### Number of early identifiers

There was no significant difference in the number of identifiers across the three conditions, *x2*(2) = 0.70, *p* = 0.705, Cramer’s V = 0.08.

#### Percent of conducive trials after the identification point

The data were not normally distributed, and a Kruskal-Wallis test using all participants’ data (identifiers and non-identifiers) found no significant difference between the conditions, *X*2(2) = 0.59, *p* = 0.746.

#### Reaction time and accuracy for all of the trials

This analysis is unaltered by the number of identifiers and non-identifiers. Therefore, the outcome is identical to what is reported in our manuscript.

## Conclusions from re-analysis

Classing the false alarms in Study 1 and 2 as identifiers made no difference to the pattern of findings that we reported in our manuscript.

1. The equivalence test results were derived from a ‘TOSTER’ spreadsheet (see Lakens, 2017) [↑](#footnote-ref-1)
2. The equivalence test results were derived from a ‘TOSTER’ spreadsheet (see Lakens, 2017) [↑](#footnote-ref-2)