2015

# Disfluent Fonts Don't Help People Solve Math Problems

Andrew Meyer
*Yale University*

Shane Frederick
*Yale University*

Terence C. Burnham
*Chapman University*, burnham@chapman.edu

Juan D. Guevera Pinto
*Louisiana State University*

Ty W. Boyer
*Georgia Southern University*

***See next page for additional authors***



## Recommended Citation

Meyer, A., S. Frederick, et al. (2015). "Disfluent fonts don't help people solve math problems," *Journal of Experimental Psychology: General* 144(2): e16. doi: 10.1037/xge0000049

# Disfluent Fonts Don't Help People Solve Math Problems



**Authors**
Andrew Meyer, Shane Frederick, Terence C. Burnham, Juan D. Guevera Pinto, Ty W. Boyer, Linden J. Ball, Gordon Pennycook, Rakefet Ackerman, Valerie A. Thompson, and Jonathon P. Schuldt

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., Pennycook, G., Ackerman, R., Thompson V., & Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General*, *144*(2), e16.

# Disfluent fonts don't help people solve math problems.

Andrew Meyer & Shane Frederick
Yale University

Terence Burnham
Chapman University

Juan D. Guevara Pinto
Louisiana State University

Ty W. Boyer
Georgia Southern University

Linden J. Ball
University of Central Lancashire

Gordon Pennycook
University of Waterloo

Rakefet Ackerman
Technion - Israel Institute of Technology

Valerie A. Thompson
University of Saskatchewan

Jonathon P. Schuldt
Cornell University

## Abstract

Prior research suggests that reducing font clarity can cause people to consider printed information more carefully. The most famous demonstration showed that participants were more likely to solve counterintuitive math problems when they were printed in hard-to-read font. However, after pooling data from that experiment with 16 attempts to replicate it, we find no effect on solution rates. We examine potential moderating variables, including cognitive ability, presentation format, and experimental setting, but we find no evidence of a disfluent font benefit under any conditions. More generally, though disfluent fonts slightly increase response times, we find little evidence that they activate analytic reasoning.

*Keywords*: fluency, disfluency, dual-system processing, reasoning, judgment.

For comments, we thank Adam Alter, Maya Bar-Hillel, Brett Buttliere, Christopher Chabris, Nicholas Epley, Elizabeth Friedman, Daniel Kahneman, Hedy Kober, Steven Malliaris, Daniel Mochon, Leif Nelson, Daniel Oppenheimer, Jennifer Savary, Joe Simmons, Daniel Simons, Uri Simonsohn, and Kathleen Vohs. For logistical support, we thank Pamela Brown, Andrew Pearlmutter, and research assistants from the Yale SOM behavioral lab.

Correspondence concerning this article can be addressed to Andrew Meyer, Marketing Department, Yale School of Management, 165 Whitney Avenue, New Haven, CT, 06511. Email: andrew.meyer@yale.edu.

Though controversy surrounds the two "systems" metaphor, most agree that thoughts differ. Many distinguish *intuitive* thoughts, released merely by exposure to stimuli, from *reflective* thoughts, occurring after deliberate deployment of additional operations (Kahneman, 2011; Shweder,1977). When intuition and reflection lead to different judgments or decisions, it is useful to understand which will prevail.

To measure the tendency toward reflection, Frederick (2005) proposed the Cognitive Reflection Test (CRT). It consists of three math problems, each with a tempting wrong answer that subsequent operations could readily disconfirm. Those who perceive the test to be more difficult score higher (Frederick, 2005; Mata, Ferreira, & Sherman, 2013), suggesting that performance on the CRT might be enhanced by cues indicating its difficulty.

Alter, Oppenheimer, Epley, and Eyre (2007) report that merely presenting the CRT in a hard-to-read ("disfluent") font provides a sufficient cue (see Figure 1). Specifically, they found that Princeton students who received the test in a disfluent font outscored those who received it in a normal font ($M_{\text{Normal}} = 1.90$, $M_{\text{Disfluent}} = 2.45$), $t_{38} = 2.25$, $p = .03$. They proposed that people who receive the test in a disfluent font misattribute the difficulty *reading* the problems to the difficulty of the problems themselves, and as a result, think more deeply.

**Figure 1: The Cognitive Reflection Test, printed in normal font (left) and disfluent font (right)**

| Normal | *Disfluent* |
|---|---|
| 1) A bat and a ball cost $1.10 in total.<br>The bat costs $1.00 more than the ball.<br>How much does the ball cost? ____ cents<br><br>2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? ____ hours<br><br>3) In a lake there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ____ days | *1) A bat and a ball cost $1.10 in total.*<br>*The bat costs $1.00 more than the ball.*<br>*How much does the ball cost?____ cents*<br><br>*2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?____ hours*<br><br>*3) In a lake there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?____ days* |

This experiment is easy to describe, and yields a result that is both surprising and potentially important. Consequently, it has received considerable attention. It is cited in over 133 academic articles and at least three bestselling books (*David and Goliath*, Gladwell, 2013, pp. 104–105; *Drunk Tank Pink: And Other Unexpected Forces That Shape How We Think, Feel, and Behave*, Alter, 2013, pp. 194–195; and *Thinking, Fast and Slow*, Kahneman, 2011, p. 65). Though a subsequent publication (Thompson et al., 2013) failed to replicate this effect in three populations, it continues to be accepted as true, and is typically cited without qualification.

To further investigate whether disfluent fonts affect CRT score, we pooled the original study from Alter et al. (2007) with all publishable replication attempts of which we were aware: three by Thompson et al. (2013) and 13 new ones that we conducted.[1] Results are summarized in Table 1 and Figure 2. Experimental details are provided in Appendix A.

[1] We posted a request to the SJDM listserv for published and unpublished replication attempts of this study. No other direct replications were reported, but several conceptual replications were. These are cited in the final paragraphs.

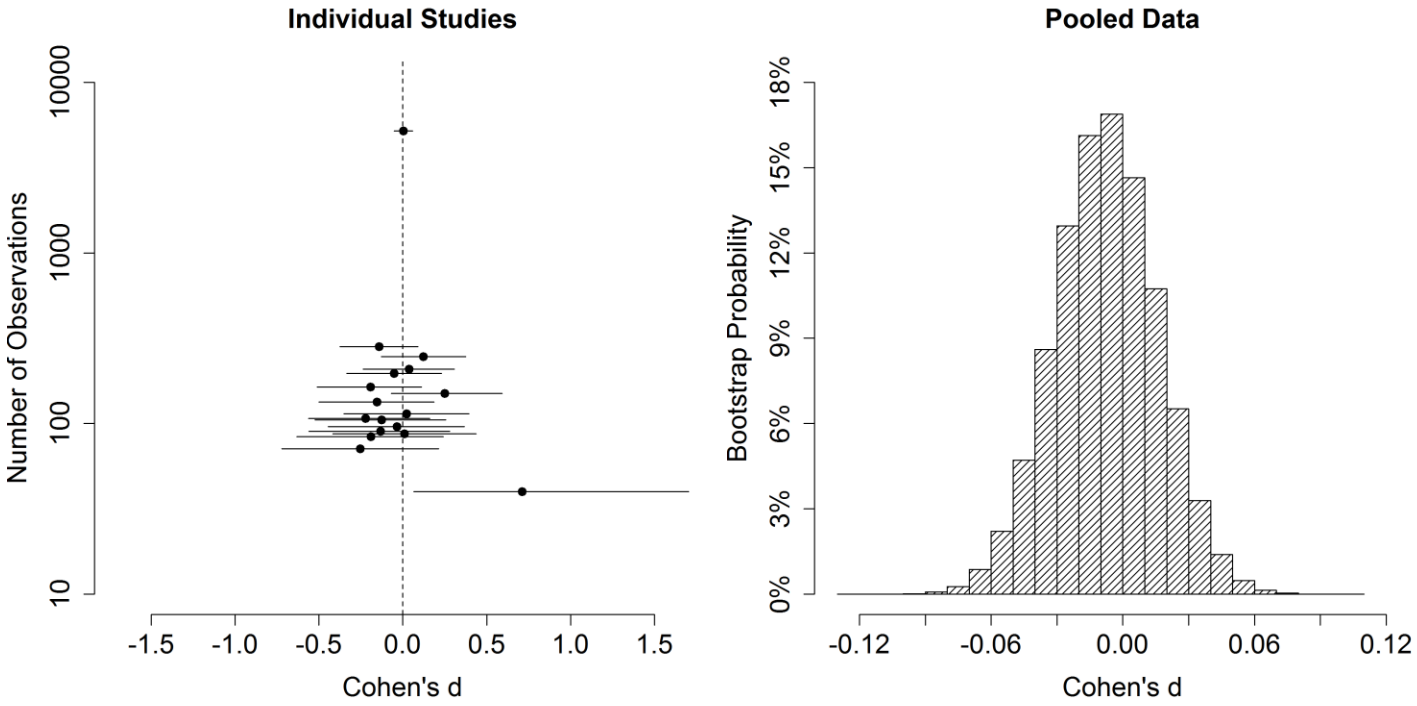
**Table 1: The effect of disfluent font on Cognitive Reflection Test scores**

| | Sample Population | Condition | # of Partici-pants | Item Solution Rates: Bat and Ball | Widgets | Lily Pads | Mean # Correct | Cohen's *d* | *p* |
|---|---|---|---|---|---|---|---|---|---|
| *1 $^{p}_{p}$ | Princeton University | Normal Font | 20 | 80 | 20 | 90 | 1.90 | 0.71 | .03 |
| | | Disfluent Font | 20 | 75 | 80 | 90 | 2.45 | | |
| *2 $^{p}_{p}$ | Technion - Israel Inst. of Tech. | גופן רגיל | 75 | 76 | 49 | 59 | 1.84 | 0.25 | .13 |
| | | גופן קשה לקריאה | 75 | 80 | 63 | 67 | 2.09 | | |
| 3 $^{p}_{p}$ | Yale University Summer Session | Normal Font | 124 | 61 | 37 | 60 | 1.58 | 0.12 | .34 |
| | | Disfluent Font | 123 | 66 | 44 | 62 | 1.72 | | |
| 4 $^{p}_{l}$ | Yale University | Normal Font | 104 | 76 | 57 | 74 | 2.07 | 0.04 | .79 |
| | | Disfluent Font | 104 | 79 | 54 | 78 | 2.11 | | |
| 5 $^{p}_{l}$ | University of Michigan | Normal Font | 56 | 52 | 29 | 50 | 1.30 | 0.02 | .91 |
| | | Disfluent Font | 58 | 53 | 28 | 52 | 1.33 | | |
| 6 $^{p}_{l}$ | New Haven Residents | Normal Font | 43 | 44 | 47 | 51 | 1.42 | 0.01 | .96 |
| | | Disfluent Font | 44 | 45 | 48 | 50 | 1.43 | | |
| 7 $^{c}_{o}$ | MTurk | Normal Font | 2,577 | 37 | 46 | 54 | 1.37 | 0.00 | .90 |
| | | Disfluent Font | 2,614 | 37 | 47 | 53 | 1.37 | | |
| 8 $^{p}_{l}$ | Chapman University | Normal Font | 48 | 63 | 56 | 65 | 1.83 | -0.04 | .86 |
| | | Disfluent Font | 48 | 52 | 65 | 63 | 1.79 | | |
| 9 $^{p}_{p}$ | Yale University | Normal Font | 99 | 81 | 68 | 79 | 2.27 | -0.05 | .72 |
| | | Disfluent Font | 98 | 80 | 63 | 80 | 2.22 | | |
| 10 $^{p}_{l}$ | Technion - Israel Inst. of Tech. | גופן רגיל | 45 | 76 | 53 | 64 | 1.93 | -0.13 | .53 |
| | | גופן קשה לקריאה | 45 | 62 | 53 | 64 | 1.80 | | |
| 11 $^{p}_{l}$ | New Haven Residents | Normal Font | 51 | 39 | 33 | 39 | 1.12 | -0.13 | .52 |
| | | Disfluent Font | 54 | 31 | 30 | 35 | 0.96 | | |
| *12 $^{c}_{o}$ | Canadians from the Internet | Normal Font | 139 | 32 | 32 | 54 | 1.17 | -0.14 | .24 |
| | | Dı$flμεητ F¤ητ | 143 | 16 | 29 | 57 | 1.02 | | |
| 13 $^{c}_{l}$ | Yale University | Normal Font | 64 | 77 | 62 | 73 | 2.12 | -0.15 | .38 |
| | | Disfluent Font | 70 | 67 | 51 | 77 | 1.96 | | |
| *14 $^{c}_{l}$ | University of Saskatchewan | Normal Font | 81 | 30 | 33 | 52 | 1.15 | -0.19 | .22 |
| | | Disfluent Font | 83 | 24 | 27 | 43 | 0.94 | | |
| 15 $^{p}_{p}$ | Yale University | Normal Font | 42 | 79 | 55 | 81 | 2.14 | -0.19 | .39 |
| | | Disfluent Font | 42 | 69 | 50 | 76 | 1.95 | | |
| 16 $^{c}_{l}$ | Georgia Southern University | Normal Font | 54 | 11 | 7 | 19 | 0.37 | -0.22 | .25 |
| | | Disfluent Font | 53 | 8 | 2 | 13 | 0.23 | | |
| 17 $^{p}_{l}$ | New Haven Residents | Normal Font | 35 | 23 | 17 | 14 | 0.54 | -0.25 | .29 |
| | | Disfluent Font | 36 | 11 | 8 | 11 | 0.31 | | |
| | Pooled | Normal Font | 3,657 | 42 | 45 | 55 | 1.43 | -0.01 | .75 |
| | | Disfluent Font | 3,710 | 41 | 46 | 55 | 1.42 | | |

*Note: Asterisks indicate previous publication. Superscripts indicate presentation format (p = paper & pencil, c = computer screen). Subscripts indicate experimental setting (p = in public, o = online, l = in lab). The last row treats each participant as one observation.*

The pooled data provide no evidence that disfluent fonts affect performance on the CRT ($M_{\text{Normal}}$ = 1.43, $M_{\text{Disfluent}}$ = 1.42), $t_{7,365} = 0.32$, $p = .75$; nor do meta-analytic techniques that treat each experiment as a single observation (Stouffer's $z = 0.72$, $p = .47$; see Rosenthal, 1978). Indeed, of the 17 experiments, only the study reported by Alter et al. (2007) finds significantly higher scores in the disfluent font condition.[2] Bootstrap resampling from the pooled data generates a 95% confidence interval ranging from 0.06 fewer to 0.05 more items answered correctly in the disfluent font conditions.[3] Manipulation checks and statistical power analyses are presented in Appendices B and C.

**Figure 2: The effect of disfluent font on Cognitive Reflection Test scores.**




*Note: The left panel of the figure graphs each individual study's sample size against its effect size. Error bars are bootstrap 95% confidence intervals. The right panel displays the bootstrap probability distribution of the effect size based on the pooled data*

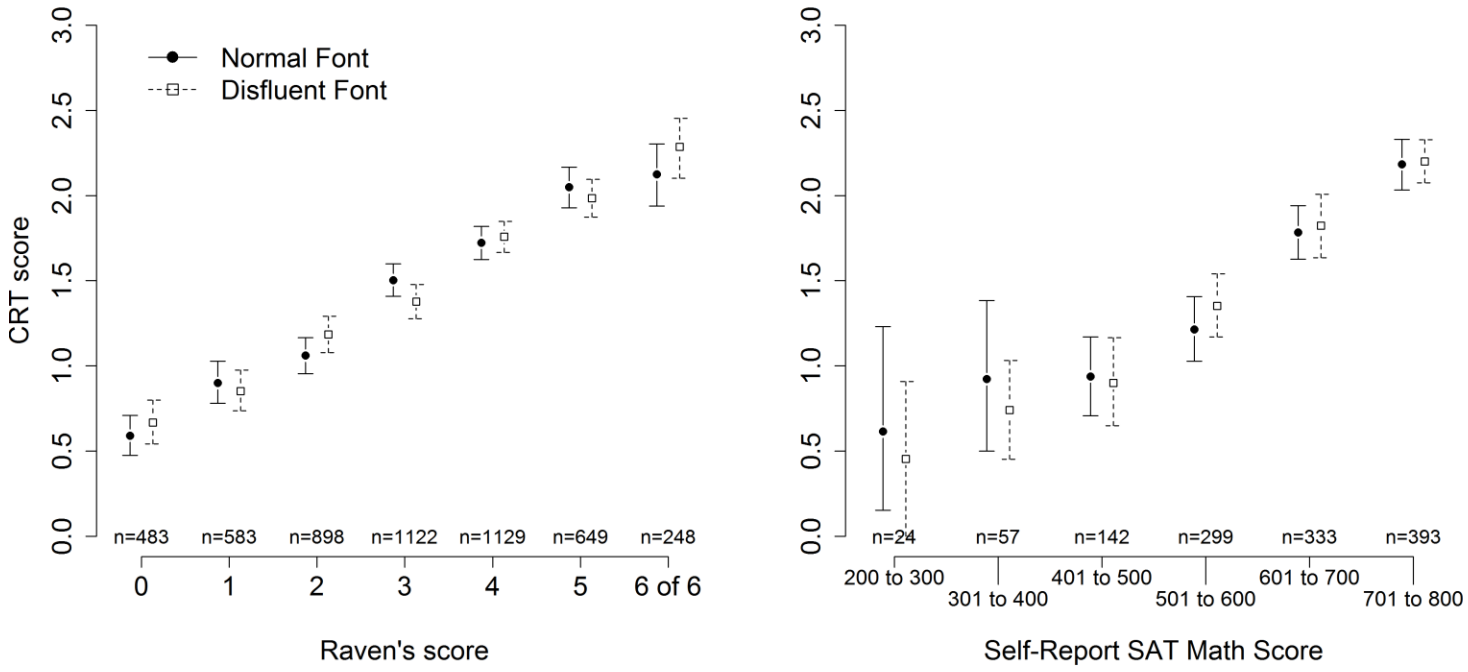
The study by Thompson and colleagues which failed to replicate the main effect of font on CRT did highlight relations between the font effect and cognitive ability. Based partly on these data, Alter, Oppenheimer, and Epley (2013) and Oppenheimer and Alter (2014) propose that the effect is restricted to those of high ability.

[2] It is often implied and sometimes claimed that disfluent fonts improve performance on the bat-and-ball problem. In fact, there was no such effect even in the original study, as shown in row 1 of Table 1. The entire effect reported in Alter et al. (2007) was driven by just one of the three CRT items: "widgets," which was answered correctly by 16 of 20 participants in the disfluent font condition, but only 4 of 20 participants in the control condition. The 20% solution rate in the control condition is poorer than every other population except Georgia Southern. It is also significantly below the 50% solution rate observed in a sample of 300 Princeton students (data available from Shane Frederick upon request). This implicates sampling variation as the reason for the original result. If participants in the control condition had solved the widgets item at the same rate as Princeton students in other samples, the original experiment would have had a $p$ value of 0.36, and none of the studies in Table 1 would exist.

[3] We supplement our parametric analyses with the empirical bootstrap because the tails of the normal distribution continue indefinitely, whereas CRT scores are censored at 0 and 3. However, both conventional parametric analysis and meta-analytic random-effect regression (see Viechtbauer, 2010) yield similar 95% confidence intervals (95% CI [0.07, 0.05] and [0.06, 0.04], respectively).

The contention that disfluent font benefits high ability participants can be tested by conducting experiments in high ability populations or by including additional measures of cognitive ability wherever the study is conducted. Neither test provides support. First, we found no effect among undergraduates at Yale University (pooling studies 4, 9, 13, and 15: $M_{\text{Normal}}$ =2.16, $M_{\text{Disfluent}}$ =2.09, $t_{621}$ = -0.81, $p$ = 0.42). Second, we found no evidence for the proposed moderation when we examined additional measures of cognitive ability. In the study conducted on MTurk (#7) and one of the studies conducted at Yale (#13), we included six items from Raven's Advanced Progressive Matrices – a widely accepted measure of general intelligence (Jensen, 1998). Disfluent font did not elevate CRT scores at any level of performance on the Raven's test, nor was there any evidence of a positive interaction ($\beta$ = -0.01, $t_{5108}$ = -0.31, $p$ = 0.76). In studies 3, 6, (part of) 7, 11, and 17 participants reported their SAT math score. We found no evidence of a positive interaction with disfluent font there either ($\beta$ = 0.0003, $t_{1244}$ = 0.70, $p$ = 0.48). Figure 3 summarizes. Furthermore, we found no significant evidence of the predicted interaction when we used educational attainment or Israeli Psychometric Entrance Test scores as proxies for cognitive ability. Details are presented in Appendices D and E.

**Figure 3: The relation between cognitive ability and the effect of disfluent font on Cognitive Reflection Test scores**




*Error bars are bootstrap 95% confidence intervals*

Alter, Oppenheimer, and Epley's (2013) proposed moderation by cognitive ability is based on the idea that disfluent fonts (or *any* manipulation that gives pause) could affect performance only to the degree that respondents can benefit from extended thought. We agree. It would be unreasonable to expect disfluent fonts to affect performance on reasoning tasks exceeding respondents' ability level, and the CRT items do, indeed, exceed this threshold for some respondents. Meyer, Spunt, and Frederick (2015) find that a substantial fraction of respondents cannot solve these problems, even when the tempting intuitive response is explicitly invalidated (e.g., when the answer blank is immediately followed by the words, "HINT: The answer is NOT 10 cents."). However, we find no effect of font even if we restrict analysis to respondents who *can* solve all the problems with such hints. In experiments #7 and #13, participants first completed the CRT, were later informed that the

intuitive lures (10 cents, 100 minutes, and 24 days) were NOT correct, and then received the opportunity to revise their answers. With the benefit of these hints, 2,145 participants (out of 5,325) managed to solve all three problems. However, the "pre-hint" CRT scores in this select group were nearly identical in the normal and disfluent font conditions ($M_{\text{Normal}} = 2.46$ vs. $M_{\text{Disfluent}} = 2.45$ items correct, $t_{2143} = -0.24$, $p = 0.81$).

The proposed moderation can even be tested in samples lacking any measure of cognitive ability aside from the CRT itself. If disfluent fonts benefit smarter participants more, they should increase variance in CRT scores – by elevating scores among those who would do well anyway, but having little effect (or even depressing) scores of those who would normally do poorly. However, our pooled data reveal no evidence of this; the variance in CRT scores is nearly identical in the two conditions ($SD^2_{\text{Normal}} = 1.45$ vs. $SD^2_{\text{Disfluent}} = 1.44$, $F_{3656, 3709} = 1.01$, $p = 0.79$).

We also tested for any effects of presentation format (paper & pencil vs. computer screen), experimental setting (in public, in lab, or online), and previous exposure to the problems. We found no evidence for moderation by these factors either (see Appendix F). We tested for potential moderators because of prior claims, reviewer requests, and curiosity. However, in light of all the data, Alter et al.'s (2007) result is not aberrant enough to motivate the search for an unobserved moderator (Simons, 2014). The distribution of effect sizes across the 17 studies is consistent with independent draws from a single unimodal distribution, as evidenced by the null result of a heterogeneity test ($I^2 = 0.03\%$ Cochran's $Q_{16} = 16.1$; $p = .44$).

Although respondents do not do any better on the CRT when it is printed in disfluent font, they do take longer to respond (Thompson et al., 2013). Three out of four studies measuring response latencies find small, but significant differences (pooling studies 7, 13, 14, and 16: Geo $M_{\text{Normal}} = 50$ seconds vs. Geo $M_{\text{Disfluent}} = 53$ seconds; $t_{5514} = 2.95$, $p = 0.003$; see Appendix G for details). These small differences might be attributed to increased reading time, or other thought processes that disfluent fonts engage, including musings about *why* the font is disfluent. It remains unclear whether this extra time implies the engagement of deeper reasoning processes. Aside from our failure to find any effect on performance, two other results weigh against this. First, Guevara Pinto (2014) manipulated CRT font fluency and found no effect on pupil dilation, which tends to reflect the engagement of effortful thought (Kahneman & Beatty, 1966; Laeng, Sirois, & Gredebäck, 2012). Second, Thompson et al. (2013) found no evidence that disfluent fonts reduced respondents' confidence in the answers they produced, as one might expect if disfluent fonts increased estimates of problem difficulty. However, it is possible that disfluent fonts engage reasoning processes that are used to justify, rather than overturn, initial intuitions. That would explain both the slowed response and the null effects on performance and confidence, though not the null effect on pupil dilation.

The experiment involving fonts and the CRT was adduced to support a more general theory of meta-cognition. But it is certainly not the only relevant datum. Disfluent fonts have been reported to improve performance on other tasks involving a conflict between intuition and reason, including belief-bias syllogisms (Experiment 4 in Alter et al., 2007), the "Moses" and "Joshua" oversight problems[4] (Song & Schwarz, 2008), and betting against a spread (Experiment 13 in Simmons & Nelson, 2006). The validity of these results should be judged on the degree to which they can be reproduced. With respect to that, we note that Alter et al. (2007) report data for just two belief-bias syllogisms and cite floor and ceiling effects to exclude four comparable items that do not show the predicted effect. One subsequent attempt to replicate an effect of font on syllogistic reasoning succeeded (Rotello & Heit, 2009), but four others have failed (Exell & Stupple, 2011; Morsanyi & Handley, 2012; Thompson et al., 2013; Trippas, Handley & Verde, 2014). Meyer and Frederick failed to replicate the effect on the Moses problem, but did replicate the effect of fonts on football bets (though the effect is tiny, and disappears altogether when the question is phrased differently). These two studies are described in Appendices H and I.

Although the more general prevalence of "desirable difficulties" (Bjork, 1994) is beyond the scope of this article, several research groups have found that disfluent fonts improve performance on memory tasks

[4] "How many animals of each kind did Moses take on the Ark?" And, "In the biblical story, what was Joshua swallowed by?" (Erickson & Mattson, 1981)

(Cotton et al, 2014; Diemand-Yauman, Oppenheimer, & Vaughan, 2011; French et al., 2013; Lee, 2013; Sungkhasettee, Friedman, & Castel, 2011; Weltman & Eakin, 2014). Though some have also failed to replicate these effects (Eitel et al., 2014; Yue, Castel, & Bjork, 2013), the balance of evidence suggests that disfluent fonts may aid memory but not reasoning – presumably because reading words more slowly benefits memory, but not reasoning.

In conclusion, after pooling across 17 experiments that manipulate the font of the CRT, we find no evidence that disfluent fonts improve performance and no support for the proposed moderation by ability. More generally, we find little evidence that disfluent fonts activate analytic reasoning.

## References

Alter, A. (2013). *Drunk tank pink: And other unexpected forces that shape how we think, feel, and behave*. Penguin.

Alter, A. L., Oppenheimer, D. M., & Epley, N. (2013). Disfluency prompts analytic thinking—But not always greater accuracy: Response to. *Cognition*,*128*(2), 252-255.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, N. (2007) Overcoming Intuition: Metacognitive Difficulty Activates Analytic Reasoning. *Journal of Experimental Psychology: General*, 136(4), 569-576

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In Metcalfe, J. E., & Shimamura, A. P. *Metacognition: Knowing about knowing*. The MIT Press.

Cotton, D., Joseph, E., Lede, M., & Ronan, D. (2014, December). *The Effect of Font Structure on Memory and Reading Time*. Poster presented at the biannual evening of psychological science, University of Connecticut.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the (): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111-115.

Eitel, A., Kühl, T., Scheiter, K., & Gerjets, P. (2014). Disfluency Meets Cognitive Load in Multimedia Learning: Does Harder-to-Read Mean Better-to-Understand? *Applied Cognitive Psychology*. 28(4), 488-501

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540-551.

Exell, R. & Stupple, E. J. N. (2011) [Text disfluency in belief-biased syllogistic reasoning]. Unpublished raw data.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, 25-42.

French, M. M. J., Blood, A., Bright, N. D., Futak, D., Grohmann, M. J., Hasthorpe, A., ... & Tabor, J. (2013). Changing fonts in education: How the benefits vary with ability and dyslexia. *The Journal of Educational Research*,*106*(4), 301-304.

Gladwell, M., (2013) *David and Goliath: Underdogs, Misfits, and the Art of Battling Giants* Little, Brown.

Guevara Pinto, J. D. (2014). Effects of perceptual fluency on reasoning and pupil dilation. (University Honors Program Thesis). Retrieved from DigitalCommons@Georgia Southern, http://digitalcommons.georgiasouthern.edu/honors-theses/2.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Strous, & Giroux.

Kahneman, D. & Beatty J. (1966) Pupil Diameter and Load on Memory. *Science*, *154*, 1583-1585.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49-81.

Laeng, B., Sirois, S., & Gredebäck, G. (2012) Pupillometry: A Window to the Preconscious? *Perspectives on Psychological Science, 7* (1), 18-27.

Lee, M. H. (2013). Effects of Disfluent Kanji Fonts on Reading Retention with E-Book. In *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on* (pp. 481-482). IEEE.

Levitt, S. D. (2004). Why are gambling markets organised so differently from financial markets? *The Economic Journal*, *114*(495), 223-246.

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: a dual-process perspective on overconfidence. *Journal of personality and social psychology*, *105*(3), 353.

Meyer, A., Spunt, R., & Frederick, S. (2015) The bat and ball problem.

Morsanyi, K., & Handley, S. J. (2012). Logic feels so good—I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 596.

Oppenheimer, D. M., & Alter, A. L. (2014). The search for moderators in disfluency research. *Applied Cognitive Psychology*.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological bulletin*, *85*(1), 185.

Rotello, C. M., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1317.

Shweder, R. A., Casagrande, J. B., Fiske, D. W., Greenstone, J. D., Heelas, P., Laboratory of Comparative Human Cognition, & Lancy, D. F. (1977). Likeness and Likelihood in Everyday Thought: Magical Thinking in Judgments About Personality [and Comments and Reply]. *Current Anthropology*, 637-658.

Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General*, *135*(3), 409.

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science*, *9*(1), 76-80.

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic bulletin & review*, *18*(5), 973-978.

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237-251.

Trippas, D., Handley, S. J., & Verde, M. F. (2014). Fluency and belief bias in deductive reasoning: New indices for old effects. *Name: Frontiers in Psychology*, *5*, 631.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48.

Weltman, D., & Eakin, M. (2014). Incorporating Unusual Fonts and Planned Mistakes in Study Materials to Increase Business Student Focus and Retention.*INFORMS Transactions on Education*, *15*(1), 156-165.

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & cognition*, *41*(2), 229-241.

## Appendix A: Detailed Methods

*Study 1:* Alter et al, (2007, p. 570) report:

> We recruited 40 Princeton University undergraduate volunteers at the student campus center to complete the three-item CRT (Frederick, 2005). Participants were seated either alone or in small groups, and the experimenter ensured that they completed the questionnaire individually. Those in the fluent condition completed a version of the CRT written in easy-to-read black Myriad Web 12-point font, whereas participants in the disfluent condition completed a version of the CRT printed in difficult-to-read 10% gray italicized Myriad Web 10-point font. Participants were randomly assigned to complete either the fluent or the disfluent version of the CRT.

*Study 2:* Thompson et al, (2013, p. 10) report:

> The three CRT problems were printed on one white page in the order they appeared in Alter et al. (2007). At the bottom of the page there were spaces for writing demographic information. There were four versions for the printed page: Two with each font type, with each one of them either including a confidence rating scale after each problem, or not. When included, confidence was rated by choosing a number on an 11-point scale marked by 0%, 10%,. . . , 100%... The participants were recruited at the campus centre (see Alter et al., 2007), grass plots, libraries, and faculty lobbies all around the campus. Each participant randomly received one of the four questionnaire versions (N > 30 in each group). They were asked to solve each problem and to indicate whether they were familiar with the problem before taking the test. Participants who knew at least one problem in advance were replaced.

*Study 3:* During the summer of 2014, six undergraduate research assistants offered summer session students in Yale University public places $1.00 to complete a survey, which consisted of the three CRT questions, printed on one side of a single sheet of paper in either normal or disfluent font (black 12-point Myriad Web vs. grey 10-point italic Myriad Web, see row 3 of table 1). A request to report SAT math score and an instruction to circle previously seen problems were printed below the CRT questions (in the same font as the questions themselves). 21% reported having seen at least one of the problems before. We included those participants. However, the result is unchanged if we exclude them (Normal = 1.45 vs. Disfluent = 1.54 items correct; $t_{192} = 0.52$, $p = .60$).

*Study 4:* During the spring semester of 2014, Yale undergraduate participants were paid $15 to complete a 45 minute packet of surveys in Yale School of Management's behavioral lab. The three CRT questions were printed on one side of a single sheet of paper in either normal or disfluent font (black 12-point Myriad Web vs. grey 10-point italic Myriad Web, see row 4 of table 1). An instruction to circle previously seen problems was printed below the questions (in the same font as the questions themselves). 45% reported having seen at least one of the problems before. We included those participants. However, the result is unchanged if we exclude them (Normal = 1.71 vs. Disfluent = 1.95 items correct; $t_{113} = 1.10$, $p = .27$).

*Study 5:* During the Fall 2009 semester, University of Michigan students from the Introductory Psychology subject pool were invited to the lab to complete an approximately 30-minute session in exchange for partial course credit. When participants indicated that they had finished a variety of computer based tasks, the experimenter presented them with the three CRT questions printed on one side of a single sheet of paper in either normal or disfluent font ((black 12-point Arial vs. black 12-point Mistral, see row 5 of table 1).

*Study 6:* During the summer of 2014, New Haven residents were paid $8 to complete a 20 minute packet of surveys including the three CRT questions, printed on one side of a single sheet of paper in either normal or disfluent font (black 12-point Myriad Web vs. grey 10-point italic Myriad Web, see row 6 of table 1). A request to report SAT math score and an instruction to circle previously seen problems were printed below the CRT questions (in the same font as the questions themselves). 39% reported having seen at least one of the problems before. We included those participants. However, the result is unchanged if we exclude them (Normal = 0.96 vs. Disfluent = 1.04 items correct; $t_{51} = 0.22$, $p = .83$).

*Study 7:* During the spring of 2014, Amazon MTurk workers were paid $1 to participate. The three CRT questions appeared on a single screen in either normal or disfluent font (black 11.5-point Arial vs. grey 8.5-

point italic Impact, see row 7 of table 1). After completing the CRT and submitting their answers, participants rated the difficulty of reading the font on a five point scale from very easy (1) to very difficult (5). Participants then received the three CRT items again, along with the hints that the answers 10 cents, 100 minutes, and 24 days were each incorrect. A prompt invited them to revise their answers. Following that, participants completed a practice item akin to those on the Raven's Advanced Progressive Matrices test (hereafter "Ravens,") which included hints of how to solve it. Six more difficult items from Raven's APM were then presented with no hints. Participants then answered some demographic questions. For a subset of those participants, the demographics included a request to report SAT math scores. Finally, participants were asked how many of the 3 CRT questions they had seen before participating that day. 57% reported having seen at least one of the problems before. We included those participants. However, the result is unchanged if we exclude them (Normal = 0.99 vs. Disfluent = 1.00 items correct; $t_{2243} = 0.19$, $p = .85$).

*Study 8:* During the spring of 2012, Chapman University students were paid at least $7 to participate in an hour of experiments at Chapman University's Economics Science Institute. The three CRT questions came at the end of the session, and were printed on one side of a single sheet of paper in either normal or disfluent font (black 12-point Times New Roman vs. black 10-point Haettenschweiler, see row 8 of table 1).

*Study 9:* During the spring semester of 2014, an undergraduate research assistant offered people in Yale University public places $1.00 to complete a survey, which consisted of the three CRT questions, printed on one side of a single sheet of paper in either normal or disfluent font (black 12-point Myriad Web vs. grey 10-point italic Myriad Web, see row 9 of table 1). An instruction to circle previously seen problems was printed below the questions (in the same font as the questions themselves). 42% reported having seen at least one of the problems before. We included those participants. Excluding them does not affect the results (Normal = 2.02 vs. Disfluent = 2.12 items correct; $t_{113} = 0.55$, $p = .58$).

*Study 10:* Participants from the Technion--Israel Institute of Technology completed a packet of surveys in a psychology lab. The three CRT questions were printed on one side of a single sheet of paper in either normal or disfluent font (see row 10 of table 1). At the bottom of the page there were spaces for writing demographic information.

*Study 11:* During the summer of 2014, New Haven residents were paid $15 to complete a 30 minute packet of surveys including the three CRT questions, printed on one side of a single sheet of paper in either normal or disfluent font (black 12-point Myriad Web vs. grey 10-point italic Myriad Web, see row 11 of table 1). An instruction to circle previously seen problems was printed below the questions (in the same font as the questions themselves). 26% had seen at least one of the problems before. We included those participants. However, the result is unchanged if we exclude them (Normal = 0.76 vs. Disfluent = 0.75 items correct; $t_{76} = -0.06$, $p = .96$).

*Study 12:* Thompson et al, (2013 page 7) report:

> …48 completed the CRT after Exp 1a; 239 were recruited from local Canadian websites (Kijiji)…

Additionally, we note that 6% of their participants reported having seen at least one of the problems before. We included those participants. Excluding them does not affect the results (Normal = 1.16 vs. Disfluent = 1.06 items correct; $t_{264} = 0.78$, $p = .44$).

*Study 13:* During the spring semester of 2014, Yale undergraduate participants were paid $15 to complete 45 minutes of computer-based surveys in Yale School of Management's behavioral lab. The three CRT questions appeared on a single screen in either normal or disfluent font (black 11.5-point Arial vs. grey 8.5-point italic Impact, see row 13 of table 1). After completing the CRT and submitting their answers, participants rated the difficulty of reading the font on a five point scale from very easy (1) to very difficult (5). Participants then received the three CRT items again, along with the hints that the answers 10 cents, 100 minutes, and 24 days

were each incorrect. A prompt invited them to revise their answers. Following that, participants completed a practice item akin to those on the Raven's Advanced Progressive Matrices test (hereafter "Ravens,") which included hints of how to solve it. Six more difficult items from Raven's APM were then presented with no hints. Participants then answered some demographic questions, and finally, were asked how many of the CRT questions they had seen before coming in to the lab that day. 59% had seen at least one of the problems before. We included those participants. However, the result is unchanged if we exclude them (Normal = 1.62 vs. Disfluent = 1.48 items correct; $t_{53} = -0.40$, $p = .69$).

*Study 14:* Thompson et al, (2013, p. 11) report:

> Participants were tested individually and were randomly assigned to the difficult and easy to read font conditions. The CRT problems were presented on a computer in a 10 point Courier New black font on a white background (easy) or a teal italicised 10 point Curlz MT font on a green background (difficult, as described in Experiment 1a). After completing each problem, participants were asked to rate their confidence on a 7-point scale with ''7'' representing the highest level of confidence. After completing the CRT, participants were administered the Shipley Institute of Living Scale, which was used to derive estimates of IQ. They also completed the Actively Open-minded Thinking Scale (AOT; Stanovich & West, 2007, 2008); this is a 41-item self-report measure of an inclination to engage in effortful vs intuitive thinking (e.g., ''No one can talk me out of something I know is right'' and ''If I think longer about a problem I will be more likely to solve it''). Participants respond on a six point scale; high scores indicate a preference for actively open-minded thinking. The time required to complete the experiment was about 30 min.

*Study 15:* During the fall semester of 2014, an undergraduate research assistant offered people in Yale University public places $3.00 to complete a survey, which consisted of the three CRT questions, printed on one side of a single sheet of paper in either normal or disfluent font (black 12-point Myriad Web vs. grey 10-point italic Myriad Web, see row 15 of table 1). An instruction to circle previously seen problems was printed below the questions (in the same font as the questions themselves). 33% reported having seen at least one of the problems before. We included those participants. Excluding them does not affect the results (Normal = 1.97 vs. Disfluent = 1.60 items correct; $t_{54} = 1.33$, $p = .19$).

*Study 16:* During the Fall 2013 and Spring 2014 semesters, Georgia Southern University undergraduate students were invited to a laboratory in the Department of Psychology to participate in a 20 minute study in exchange for course credit. Participants viewed the three CRT questions in either a normal or a hard-to-read font (black 16-point Myriad Web vs. grey 8-point italic Myriad Web, see row 16 of table 1) on a Tobii TX-300 eye-tracking system, randomly intermixed with four additional reasoning problems, interleaved with non-demanding demographic questions that were included to provide baseline occulomotor measures.

*Study 17:* During the summer of 2014, New Haven residents were paid $6 to complete a 15 minute packet of surveys including the three CRT questions, printed on one side of a single sheet of paper in either normal or disfluent font (black 12-point Myriad Web vs. grey 10-point italic Myriad Web, see row 17 of table 1). An instruction to circle previously seen problems was printed below the questions (in the same font as the questions themselves). 15% had seen at least one of the problems before. We included those participants. However, if we exclude them, we actually find a significant disfluent font detriment ($M_{\text{Normal}} = 0.41$ vs. $M_{\text{Disfluent}} = 0.04$ items correct; $t_{60} = -2.19$, $p = 0.03$).

## Appendix B: Manipulation Checks

*Study 1:* We quote from page 570 of Alter et al. (2007):

> A separate sample of 13 participants rated (on a 5-point scale) the disfluent font ($M$ = 3.08, $SD$ = 0.76) as being more difficult to read than the fluent font ($M$ = 1.54, $SD$ = 0.87), $t(12) = 3.55$, $p < .01$, $\eta^2 = .51$.

*Study 2:* We quote from page 10 of Thompson et al. (2013):

> A pre-test was used to choose the fluent and disfluent fonts for the study. Twenty participants rated the legibility of one base font and four other font types on a five-point scale ranging from 1 (illegible) to 5 (easier to read than the base font). The chosen fluent font was identified as easy to read (i.e., rated 4 or higher) by all participants. The chosen disfluent font was rated as illegible (1) by two participants, as legible with effort (2) by fourteen participants, and as legible but cause feeling of discomfort (3) by four participants. No participant characterised this font as easy to read (4) or easier than the regular font (5).

*Study 3:* None

*Study 4:* None

*Study 5:* None.

*Study 6:* None.

*Study 7:* After completing the CRT and submitting their answers, participants rated (on a 5-point scale) the disfluent font as being more difficult to read than the normal font, $M$s = 3.7 vs. 1.8, $t_{4989} = 58.3$, $p < .001$, $\eta^2 = .41$.

*Study 8:* None.

*Study 9:* None.

*Study 10:* Same font and population as study 2.

*Study 11:* None.

*Study 12:* We quote from page 3 of Thompson et al. (2013):

> This combination was chosen on the basis of a pilot study that showed it to be particularly difficult to read but still legible.

*Study 13:* After completing the CRT and submitting their answers, participants rated (on a 5-point scale) the disfluent font as being more difficult to read than the normal font, $M$s = 3.7 vs. 2.0, $t(116) = 9.71$, $p < .001$, $\eta^2 = .45$.

*Study 14:* None.

*Study 15:* Same font as study 1.

*Study 16:* None.

*Study 17:* Same font as study 1.

## Appendix C: Power Analyses

NOTE: The computations below assume conventional levels of statistical significance ($\alpha = .05$):

*Main effect of font:* Assuming an effect as large as the one reported by Alter et al. (2007), Cohen's *d* = .71, we'd be essentially certain to detect it in our pooled data set.[5] Indeed, even assuming an effect one fourth as large as the one reported by Alter et al. (2007), our pooled data would detect it 99.9999993% of the time.

*Main effect of font (among participants who could solve all three problems with hints):* Assuming the effect size reported by Alter et al. (2007), Cohen's *d* = .71, our statistical power is essentially 100%. Indeed, even if the true effect were one fourth as large, our pooled data would detect it 98.4% of the time.

## Appendix D: Raven's Matrices

Each Raven's item consists of a three by three matrix with one missing element. Participants must select which of eight presented options best completes the pattern. We used items # 2, 8, 14, 20, 26, and 34 from the second set of the Raven's Advanced Progressive Matrices. Point biserial correlations between items are presented in table A1. The six items form a scale with a standardized alpha of .66. Table A2 presents individual item solution rates for Yale University undergrads, and MTurk workers, separately.

**Table A1: inter-item correlations (bottom triangle) and numbers of observations (top triangle)**

| | Item #2 | Item #8 | Item #14 | Item #20 | Item #26 | Item #34 |
|---|---|---|---|---|---|---|
| Item # 2 | -- | 5,109 | 5,102 | 5,091 | 5,090 | 5,085 |
| Item # 8 | .44 | -- | 5,087 | 5,077 | 5,075 | 5,070 |
| Item #14 | .39 | .38 | -- | 5,079 | 5,077 | 5,072 |
| Item #20 | .19 | .20 | .25 | -- | 5,074 | 5,069 |
| Item #26 | .20 | .22 | .24 | .15 | -- | 5,074 |
| Item #34 | .18 | .19 | .23 | .17 | .19 | -- |

**Table A2: solution rates**

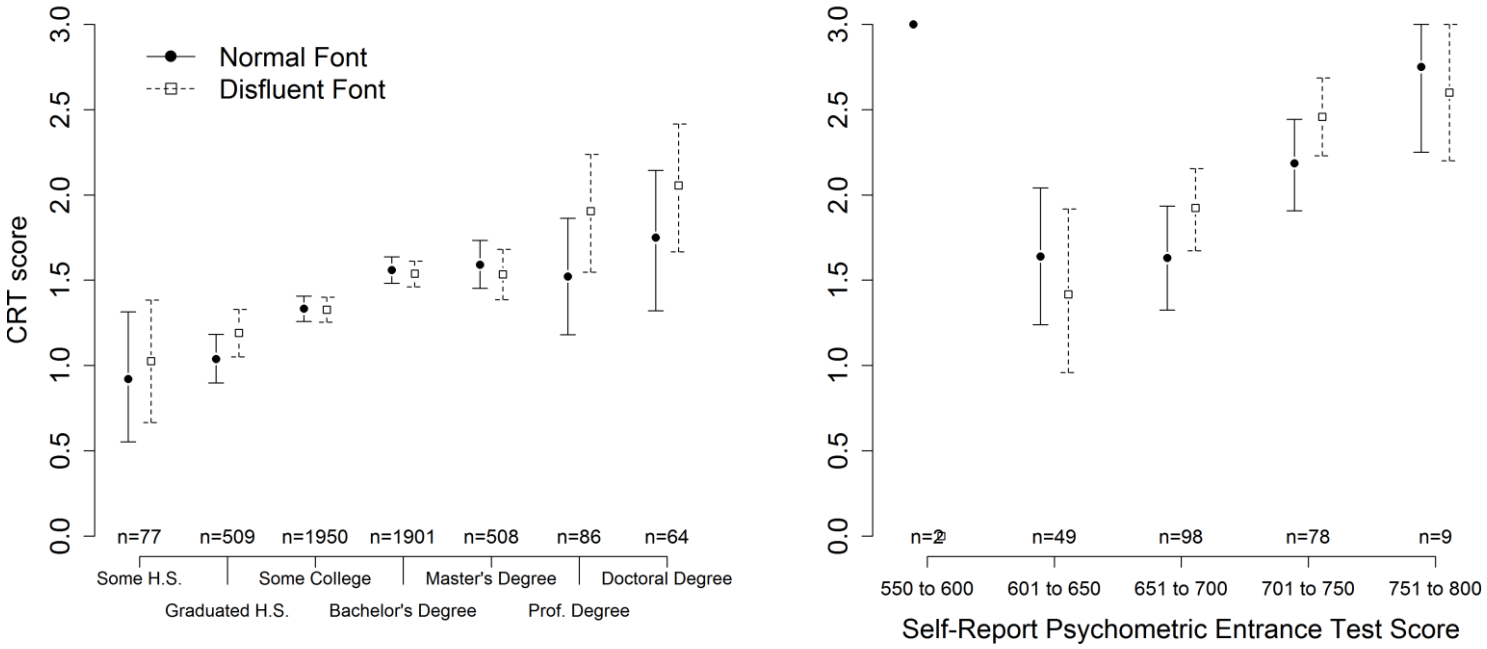
| | Yale | MTurk |
|---|---|---|
| Item # 2 | 96% | 79% |
| Item # 8 | 95% | 70% |
| Item #14 | 85% | 59% |
| Item #20 | 63% | 38% |
| Item #26 | 45% | 31% |
| Item #34 | 29% | 16% |

[5] Given the machine epsilon, our statistical power is indistinguishable from 100%.

## Appendix E: Relation with self-report educational attainment and PET score

In the study conducted on mTurk (#7) and one of the studies conducted at Yale (#13), participants reported their highest educational attainment. We see no evidence of a positive interaction between educational attainment and the effect of disfluent font ($\beta$ = -0.01, $t_{5091}$ = -0.24, $p$ = 0.81). Participants in both of the experiments conducted at the Technion (#2 and #9) reported scores on Israel's Psychometric Entrance Test. We see no significant evidence of a positive interaction between PET score and the effect of disfluent font (PET: $\beta$ = 0.005, $t_{232}$ = 1.50, $p$ = 0.14). See figure A1.

**Figure A1: The relation between cognitive ability and the effect of disfluent font on CRT scores**




*Error bars are bootstrap 95% confidence intervals.*

## Appendix F: Other potential moderators

Table A3 presents results separately for each presentation format (paper & pencil vs. computer screen). It shows no font effect in either format.

**Table A3: The effect of disfluent font on CRT scores by presentation format**

| Presentation Format | Condition | # of Partici-pants | Item Solution Rates: Bat and Ball | Item Solution Rates: Widgets | Item Solution Rates: Lily Pads | Mean # Correct | Cohen's *d* and 95% CI | *p* |
|---|---|---|---|---|---|---|---|---|
| Paper & pencil (Exp. #s 1, 2, 3 4, 5, 6, 8, 9, 10,11, 15, and 17) | Normal Font | 742 | 65 | 47 | 62 | 1.73 | 0.01 -0.09 : 0.11 | .82 |
| | Disfluent Font | 747 | 63 | 49 | 63 | 1.75 | | |
| Computer screen (Exp. #s 7 and 12, 13, 14, and 16) | Normal Font | 2,915 | 37 | 45 | 54 | 1.35 | -0.01 -0.06 : 0.04 | .65 |
| | Disfluent Font | 2,963 | 36 | 45 | 53 | 1.34 | | |

Table A4 presents results separately for each experimental setting (in public, online, or in lab). It shows no font effect in any setting.[6]

**Table A4: The effect of disfluent font on CRT scores by experimental setting**

| Experimental Setting | Condition | # of Partici-pants | Item Solution Rates: Bat and Ball | Widgets | Lily Pads | Mean # Correct | Cohen's *d* and 95% CI | *p* |
|---|---|---|---|---|---|---|---|---|
| In public (Exp. #s 1, 2, 3, 9, and 15) | Normal Font | 360 | 73 | 49 | 69 | 1.91 | 0.09 -0.06 : 0.23 | .23 |
| | Disfluent Font | 358 | 73 | 56 | 71 | 2.00 | | |
| Online (Exp. #s 7 and 12) | Normal Font | 2,716 | 36 | 45 | 54 | 1.36 | -0.00 -0.06 : 0.05 | .90 |
| | Disfluent Font | 2,757 | 36 | 46 | 53 | 1.35 | | |
| In lab (Exp. #s 4, 5, 6, 8, 10, 11, 13, 14, 16, and 17) | Normal Font | 581 | 51 | 41 | 54 | 1.46 | -0.07 -0.19 : 0.04 | .21 |
| | Disfluent Font | 595 | 47 | 38 | 52 | 1.37 | | |

In the experiments conducted at Yale (#s 3, 4, 9, 13, and 15), on MTurk (#7), and among New Haven residents (#s 6, 11, and 17), we asked participants whether they had seen any of the problems before. In those experiments, many said they had (53%). It isn't obvious how prior exposure to the problems should interact with a font fluency manipulation. On one hand, the efficacy of a font manipulation might be reduced if respondents replace or supplement reasoning with recollection. On the other hand, prior exposure might increase fluency, allowing for a larger potential impact of manipulations which decrease it.

Table A5 presents results separately for virgin respondents and those who report having seen at least one CRT item previously. It shows no font effect on either type.

**Table A5: The effect of disfluent font on CRT scores by self-report previous exposure**

| Self-report previous exposure | Condition | # of Partici-pants | Item Solution Rates: Bat and Ball | Widgets | Lily Pads | Mean # Correct | Cohen's *d* and 95% CI | *p* |
|---|---|---|---|---|---|---|---|---|
| Never seen before | Normal Font | 1,986 | 35 | 34 | 47 | 1.17 | -0.00 -0.06 : 0.06 | .97 |
| | Disfluent Font | 2,012 | 34 | 36 | 47 | 1.17 | | |
| Seen before | Normal Font | 1,671 | 50 | 58 | 66 | 1.74 | -0.02 -0.08 : 0.05 | .63 |
| | Disfluent Font | 1,698 | 50 | 58 | 64 | 1.72 | | |

Further, after excluding participants who report having seen the problems before, the data still show no evidence of moderation by intelligence. There is no effect among Yale students ($M_{\text{Normal}} = 1.85$ vs. $M_{\text{Disfluent}} = 1.88$, $t_{339} = 0.25$, $p = 0.80$), no interaction between disfluent fonts and Raven's score ($\beta = -0.02$, $t_{2083} = -0.73$, $p = 0.47$), no interaction between disfluent fonts and SAT scores ($\beta = 0.0003$, $t_{603} = 0.52$, $p = 0.60$) , no interaction between disfluent fonts and educational attainment ($\beta = -0.04$, $t_{2070} = -0.75$, $p = 0.45$), and no interaction between disfluent fonts and PET scores ($\beta = 0.005$, $t_{232} = 1.50$, $p = 0.14$). There is no effect of disfluent fonts on CRT performance among participants who got 3 out of 3 with the benefit of the hints ($M_{\text{Normal}}$

[6] However, at one point in our data collection, the "in public" experiments (which were conducted on participants who were stopped on the spot and asked to participate) hinted at superior performance among those who completed the disfluent version (pooled *p* equaled 0.11). We pursued this suggestion and ran one additional "in public" experiment (#15), but it did not lend further support for any positive effect of disfluent fonts (the pooled p increased to 0.23). We note that this experimental setting is susceptible to a failure of random assignment – and, thus, to misinterpreting selection effects as treatment effects. Whenever respondents can drop out following their inspection of the survey materials, disfluent fonts might cause the least motivated to do so, thereby improving the average "quality" of the respondents who remain.

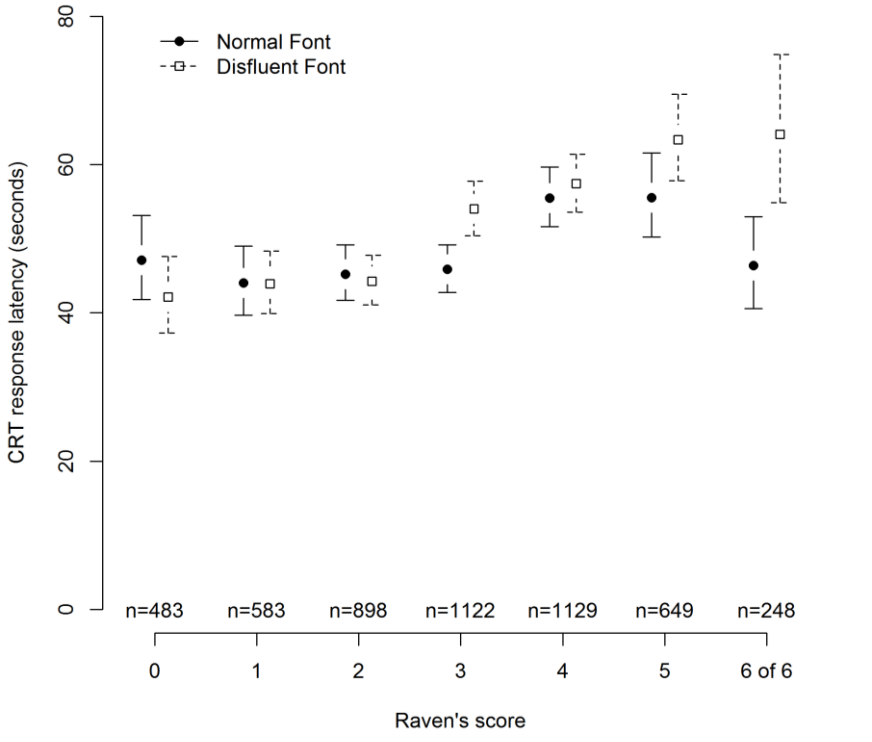
= 2.27 vs. $M_{Disfluent}$ = 2.22 items correct, $t_{658}$ = -0.66, $p$ = 0.51), and no effect of disfluent fonts on the variance in CRT scores ($SD^2_{Normal}$ = 1.34 vs. $SD^2_{Disfluent}$ = 1.32, $F_{1985, 2011}$ = 1.01, $p$ = 0.76).

Excluding previously exposed participants reveals no evidence of moderation by anything else either. Among virgin participants, there is no effect in either presentation format: neither paper & pencil ($M_{Normal}$ = 1.56 vs. $M_{Disfluent}$ = 1.61, $t_{1159}$ = 0.86, $p$ = 0.39) nor computer screen ($M_{Normal}$ = 1.01 vs. $M_{Disfluent}$ = 0.99, $t_{2835}$ = 0.49, $p$ = 0.62). There is also no effect in any experimental setting: neither in public ($M_{Normal}$ = 1.76 vs. $M_{Disfluent}$ =1.89, $t_{553}$ = 1.40, $p$ = 0.16), nor online ($M_{Normal}$ = 1.01 vs. $M_{Disfluent}$ = 1.01, $t_{2509}$ = -0.06, $p$ = 0.95), nor in lab ($M_{Normal}$ = 1.22 vs. $M_{Disfluent}$ = 1.17, $t_{930}$ = 0.63, $p$ = 0.53). Finally, a heterogeneity test still shows no significant evidence of an unobserved moderator ($I^2$ = 13.36%, Cochran's $Q_{16}$ = 20.4, $p$ = 0.20).

## Appendix G: Response latency

Pooling across the four studies that measured response latency (#s 7, 13, 14, and 16) shows that, overall, participants took slightly longer to respond when the questions were printed in disfluent font (Geo $M_{Normal}$ = 50 vs. Geo $M_{Disfluent}$ = 53 seconds; $t_{5514}$ = 2.95, $p$ = 0.003). However, the overall effect is completely driven by the most intelligent participants; only the most intelligent participants spent extra time in the disfluent font condition (Interaction: $\beta$ = 0.05, $t_{5108}$ = 3.34, $p$ < .001). We are unsure how to interpret this.

**Figure A2: The relation between Raven's score and the effect of disfluent font on Cognitive Reflection Test response latency**




*Error bars are bootstrap 95% confidence intervals around the geometric mean.*

## Appendix H: The Moses illusion

We included the Moses problem (Erickson & Mattson, 1981) as the third question in a packet of surveys administered to 539 people waiting to watch 4th of July fireworks on the Boston Esplanade in 2012. It was either printed in normal font or disfluent font, as printed below. Unlike Song and Schwarz (2008), we did not alert participants to be on the lookout for malformed questions.

3. How many animals of each kind did Moses take on the Ark? _____

*3. How many animals of each kind did Moses take on the Ark? _____*

The intuitive error is "2", the number of animals of each kind that *Noah* is supposed to have taken on the Ark. The correct answer is ambiguous, but "0" is not wrong. We found no significant effect on the percentage responding "2" (Normal = 69% vs. Disfluent = 66%, $z = 0.80$, $p = 0.43$), nor on the percentage responding "0" (Normal = 17% vs. Disfluent = 16%, $z = -0.48$, $p = 0.63$).

## Appendix 9: Betting the favorite to cover the spread

Levitt (2004) finds that most gamblers bet the favorite to cover the spread, despite the fact that, historically, favorites are no more likely to cover the spread than underdogs. Simmons & Nelson (2006) explain this by suggesting that bettors perform an attribute substitution (Kahneman & Frederick, 2002), in which they unwittingly substitute an easier question (whether the favorite will beat the underdog) for the relevant question (whether the favorite will beat the underdog by the specified spread).

In their 13th experiment, Simmons & Nelson (2006) observe that the favorite bias is attenuated by printing the question in disfluent font. We attempted to replicate this result. Before each Thursday night game during the first 13 weeks of the 2014 NFL season, we used MTurk to conduct one or more betting experiments in which workers were randomly assigned to a normal or disfluent font condition and asked to bet against the current Las Vegas spread. In some experiments, we asked participants whether or not they would bet on the favorite to cover the spread. In others, we asked whether or not they would bet on the underdog to cover. And in others, we asked whether they would bet on the favorite to cover or bet on the underdog to cover. Table A6 reports the percentage betting the favorite (or refusing the underdog) in each font condition, for each experiment.

**Table A6: The effect of disfluent font on betting the favorite to cover the spread**

| Question Format | Favorite | Underdog | Spread | Font | # of Parti-cipants | % Betting the Favorite | Cohen's *d* | *p* |
|---|---|---|---|---|---|---|---|---|
| Fav? Y/N | Dolphins | Bills | 5.5 | Normal Font<br>*Disfluent Font* | 245<br>257 | 60<br>52 | 0.17 | .06 |
| Fav? Y/N | Saints | Panthers | 2.5 | Normal Font<br>*Disfluent Font* | 254<br>246 | 74<br>67 | 0.15 | .09 |
| Fav? Y/N | Chiefs | Raiders | 7.5 | Normal Font<br>*Disfluent Font* | 216<br>195 | 73<br>68 | 0.11 | .27 |
| Fav? Y/N | Patriots | Jets | 9.5 | Normal Font<br>*Disfluent Font* | 286<br>309 | 68<br>64 | 0.09 | .26 |
| Which? F/U | Broncos | Chargers | 8.0 | Normal Font<br>*Disfluent Font* | 501<br>516 | 70<br>66 | 0.09 | .14 |
| Und? Y/N | Dolphins | Bills | 5.5 | Normal Font<br>*Disfluent Font* | 244<br>249 | 48<br>43 | 0.08 | .35 |
| Fav? Y/N | Broncos | Chargers | 8.0 | Normal Font<br>*Disfluent Font* | 302<br>302 | 66<br>62 | 0.08 | .35 |
| Which? F/U | Packers | Vikings | 8.0 | Normal Font<br>*Disfluent Font* | 327<br>294 | 68<br>66 | 0.05 | .56 |
| Fav? Y/N | Seahawks | Packers | 5.5 | Normal Font<br>*Disfluent Font* | 311<br>297 | 63<br>62 | 0.02 | .79 |
| Und? Y/N | Lions | Bears | 7.0 | Normal Font<br>*Disfluent Font* | 215<br>250 | 43<br>42 | 0.02 | .86 |
| Fav? Y/N | Bengals | Browns | 7.0 | Normal Font<br>*Disfluent Font* | 231<br>215 | 54<br>53 | 0.01 | .89 |
| Fav? Y/N | Falcons | Buccaneers | 7.0 | Normal Font<br>*Disfluent Font* | 307<br>301 | 64<br>63 | 0.01 | .92 |
| Fav? Y/N | Ravens | Steelers | 2.5 | Normal Font<br>*Disfluent Font* | 300<br>312 | 47<br>46 | 0.00 | .96 |
| Fav? Y/N | Lions | Bears | 7.0 | Normal Font<br>*Disfluent Font* | 253<br>214 | 53<br>53 | 0.00 | .98 |
| Fav? Y/N | Redskins | Giants | 4.0 | Normal Font<br>*Disfluent Font* | 320<br>298 | 56<br>56 | -0.01 | .91 |
| Which? F/U | Bengals | Browns | 7.0 | Normal Font<br>*Disfluent Font* | 388<br>388 | 51<br>52 | -0.01 | .89 |
| Which? F/U | Colts | Texans | 2.5 | Normal Font<br>*Disfluent Font* | 626<br>603 | 70<br>71 | -0.03 | .61 |
| Which? F/U | Saints | Panthers | 2.5 | Normal Font<br>*Disfluent Font* | 434<br>414 | 63<br>64 | -0.04 | .58 |
| Und? Y/N | Chiefs | Raiders | 7.5 | Normal Font<br>*Disfluent Font* | 215<br>192 | 51<br>53 | -0.04 | .70 |
| Which? F/U | Patriots | Jets | 9.5 | Normal Font<br>*Disfluent Font* | 295<br>297 | 69<br>71 | -0.06 | .50 |
| | | | Pooled | Normal<br>*Disfluent Font* | 6,270<br>6,179 | 62<br>60 | 0.03 | .07 |

Our results are consistent with those reported by Simmons & Nelson (2006); although the effect is small, fewer people bet on the favorite when the problem is printed in disfluent font (Normal = 61.5% favorite vs. Disfluent = 60.0% favorite, $z_{12,417} = 1.79$, $p = 0.073$). This effect is consistent both with Simmons & Nelson's (2006) explanation (that disfluency increases reflection, which decreases a favorite bias) and with an alternative: that more people respond randomly when the question is hard to read. However, the effect is at least partially mediated by slower response (Mediation Effect = -0.004; 95% CI: -0.006 to -0.003; $p < 0.001$), which seems to support the Simmons and Nelson interpretation.

Disfluent fonts increased underdog betting only in experiments in which a "yes" response indicated a bet on the favorite (Normal = 61.4% favorite vs. Disfluent = 58.8% favorite, $z_{5969} = 2.05$, $p = 0.041$). There was no significant font effect in experiments in which a "yes" response indicated a bet on the *underdog* (Normal = 47.0% favorite vs. Disfluent = 45.4% favorite, $z_{1363} = 0.59$, $p = 0.56$), nor in experiments in which participants reported their betting decision by selecting either the favorite team or the underdog team (Normal = 65.5% favorite vs. Disfluent = 65.4% favorite, $z_{5081} = 0.13$, $p = 0.90$). Thus, though we could replicate a small effect using a similar procedure, the effect is not robust and is not straightforwardly explained in terms of overriding an intuition to bet on the favorite.