Judgment and Decision Making, Vol. 13, No. 3, May 2018, pp. 246–259

The non-effects of repeated exposure to the Cognitive Reflection Test

Andrew Meyer* Elizabeth Zhou† Shane Frederick‡

Abstract

We estimate the effects of repeated exposure to the Cognitive Reflection Test (CRT) by examining 14,053 MTurk subjects who took the test up to 25 times. In contrast with inferences drawn from self-reported prior exposure to the CRT, we find that prior exposure usually fails to improve scores. On average, respondents get only 0.024 additional items correct per exposure, and this small increase is driven entirely by the minority of subjects who continue to spend time reflecting on the items. Moreover, later scores retain the predictive validity of earlier scores, even when they differ, because initial success and later improvement appear to measure the same thing.

Keywords: Cognitive Reflection Test, repeated testing

1 Introduction

The Cognitive Reflection Test (below) is intended to measure the disposition or ability to engage in reflective thought (Frederick, 2005), as it requires, among other things, that respondents override intuitively appealing but incorrect answers. The test has become popular because it is easy to administer, maps onto the central distinction underlying many dual process theories (Kahneman & Frederick, 2002; Evans & Stanovich, 2013), and predicts things that people care about, such as patience (Frederick, 2005; Shenhav, Rand & Greene, 2017), risk tolerance (Frederick, 2005; Campitelli & Labollita, 2010), willingness to admit ignorance (Fernbach et al., 2012), ability to differentiate real news from fake news (Pennycook & Rand, 2017), and religiosity (Pennycook et al., 2012; Shenhav, Rand & Greene, 2012).

A bat and a ball cost $110 in total. The bat costs $100 more than the ball. How much does the ball cost? _____ dollars

If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ mins

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 lake, how long would it take for the patch to cover half the lake? _____ days

Since the test has become popular, frequent subjects in psychological studies (e.g., MTurkers, some undergraduates, etc.) may encounter it multiple times. Although respondents usually receive no feedback, solutions are readily available online (there are currently over 300 YouTube videos explaining how to solve the bat & ball problem). This paper investigates the effect of repeated exposure on scores and on the predictive validity of those scores, by tracking the performance of 14,053 MTurkers who took the test from 1 to 25 times between November, 2013 and April, 2015. Table 1 partitions the data into four series of surveys and provides an overview.

Four results are notable: (1) self-reports of prior exposure markedly exaggerate the effect of prior exposure on score. (2) The average effect of prior exposure is small. (3) This small average effect is driven almost entirely by the subset of subjects who continue to spend time on the test. (4) The test's predictive validity is robust to prior exposure, in part because subsequent scores are an excellent proxy for initial scores, and in part because initial performance and later improvement both diagnose the tendency to reflect.

The observation that more active MTurkers perform better on the CRT (Chandler et al., 2013) has sparked worries that prior exposure may invalidate the test. In response, researchers have asked subjects whether they've seen the test before (Haigh, 2016; Stieger & Reips, 2016), which items they've seen before (Haigh, 2016), or how many of the three items they've seen before (Thomson & Oppenheimer, 2016, and us, throughout our Fall 2014 series). In all cases, respondents who report having seen the test before do better – often by a lot, as shown in the middle column of Table 2.

The relation between reported exposure and performance is usually interpreted as an effect of exposure. However, that causal inference requires at least two assumptions: first, that mathematical ability is uncorrelated with the degree of exposure, and second, that mathematical ability is uncorrelated with the ability to recall exposure. The rightmost column of Table 2 shows that the second assumption is badly violated.

For comments, we thank Maya Bar-Hillel, Jonathan Baron, Eric Bradlow, Chris Chabris, Zoe Chance, Lee Follis, Alex Fulmer, Reid Hastie, Ryan Hauser, Dan Kahan, Daniel Kahneman, Jin Kim, Amanda Levis, Steve Malliaris, Hillary Parent, Kariyushi Rao, Taly Reich and Daniel Read.

Results reported here are supported by those reported in this issue by Stagnaro, Pennycook & Rand (2018).

*Booth School of Business, University of Chicago. Email: Andrew.Meyer@chicagobooth.edu.

†Undergraduate student, Massachusets Institute of Technology.

‡Yale School of Management, Yale University.

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

TABLE 1: Data overview

			Number of . . .			# of subjects previously appearing in. . .		
Series	Dates	Measures	Surveys	Obs.	Subjects	Fall 2013	Spring 2014	Fall 2014
Fall 2013	11/19/2013 to 11/21/2013	CRT	2	1,020	982			
Spring 2014	3/14/2014 to 6/6/2014	CRT with Feedback, Ravens, Linda	6	6,843	5,191	327		
Fall 2014	9/3/2014 to 1/12/2015	CRT, SAT, Self-reported exposure	17	14,500	6,910	243	1,298	
Winter 2015	1/19/2015 to 4//9/2015	Modified CRT	5	6,125	4,670	204	851	1,610

TABLE 2: CRT scores by self-reported exposure # of scores

# of items subject *reports* having seen before	First time subject appears in our data (earlier exposure unknown)	Returning subjects (Have definitely seen all 3 items at least once before)
0	1.02 2339	1.10 616
1	1.61 610	1.42 327
2	1.56 414	1.56 425
3	1.64 1743	1.97 6649

As one might have predicted from the general tendency for mental abilities to correlate positively (Jensen, 1998; Lubinski & Humphreys, 1990; Unsworth, 2010), the ability to recall exposure to these problems is strongly correlated with the ability to solve them. Thus, self-reported prior exposure would diagnose superior performance (identifying those who are good at these problems) even if actual exposure had no effect.[1] (For more, see Appendix A.)

Table 3 shows that the first assumption may be violated as well. It sorts subjects by the number of times they appear, and reveals that more frequent subjects have higher CRT scores, even on their first trial, suggesting that mathematically inclined subjects expose themselves to such tasks more frequently and, correspondingly, are more likely to have had prior exposure to the CRT. (For more, see Appendix B.)

The best way to asess the effect of exposure, per se, is to track performance of the same subjects over time. Although we don't know the exposure histories of people entering our study, we can track subjects who appear multiple times during the Fall of 2014. These longitudinal effects are revealed in Table 3 as changes in the numbers moving down any column. They show a small effect of exposure (scores rise slightly) and a large effect on response latencies (subjects are spending much less time on the test). Scores improve by an average of only 0.024 items per exposure – a tiny fraction of the 0.829 item improvement implied by self-reports.[2]

Many have expressed concerns that the CRT will be destroyed by its popularity (Chandler et al., 2013; Baron et al., 2015; Haigh, 2016; Stieger & Reips, 2016; Thomson & Oppenheimer, 2016). The most common worry is that respondents will learn all the answers, eliminating any variance, and, hence, any covariance with other constructs of interest. But this concern is overhyped. Though a rise in scores reduces variance in elite populations, for which ceiling effects are already a problem (e.g., Princeton undergraduates), it increases variance in less elite populations, for which floor effects are the current problem. MTurkers are likely the most heavily exposed population (Rand et al., 2014), yet plenty of variance remains.

The concern that the CRT items "will lose some of their predictive power through repeated use" (Baron et al., 2015, page 268) reflects not only the worry about ceiling effects, but also the worry that the ability to learn the correct answers may measure something different from the ability to solve the problems in the first place. Among subjects who take the CRT multiple times, one can model current score (S_n) as initial score (S_1) plus the improvement afforded by further opportunities to reflect ($R_{2:n}$), plus an error term (ε_n), to capture changes in score that are uncorrelated with reflection:

$$S_n = S_1 + R_{2:n} + \varepsilon_n$$

From this perspective, the predictive validity of current score will remain intact if it closely resembles the initial score ($R_{2:n}$

[1] In addition to the selection forces we describe, reverse causation is also possible. For example, problems that are easy to solve may feel more familiar, and participants experiencing persistent difficulties may explain them away by invoking problem novelty.

[2] We estimate the repeat exposure effect by regressing CRT score against number of previous exposures with a non-parametric control for total number of appearances in the data. We estimate the self-report "effect" by regressing CRT score against percentage of items *reported* seen before. Both regressions are ordinary least squares.

The modest improvement across successive trials within our study likely *exaggerates* the effect of repeated exposure to the CRT, because some of these subjects probably encountered it in other studies between their nth and n+1st exposures in *our* study.

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

TABLE 3: Mean CRT scores and geometric mean seconds to respond across repeated testing

nth appearance in series	# of times respondent appeared in Fall 2014 series (# of respondents)						
	1 (3992)	2 (1348)	3 (614)	4 (315)	5 (183)	6 (129)	7+ (329)
1st	1.41 (52)	1.44 (43)	1.48 (38)	1.60 (38)	1.61 (36)	1.77 (30)	1.76 (27)
2nd		1.52 (27)	1.52 (23)	1.66 (23)	1.64 (21)	1.78 (19)	1.80 (18)
3rd			1.61 (19)	1.70 (19)	1.70 (18)	1.77 (15)	1.80 (15)
4th				1.75 (17)	1.75 (15)	1.82 (13)	1.81 (13)
5th					1.76 (13)	1.88 (12)	1.84 (12)
6th						1.95 (13)	1.89 (12)
7th							1.91 (11)

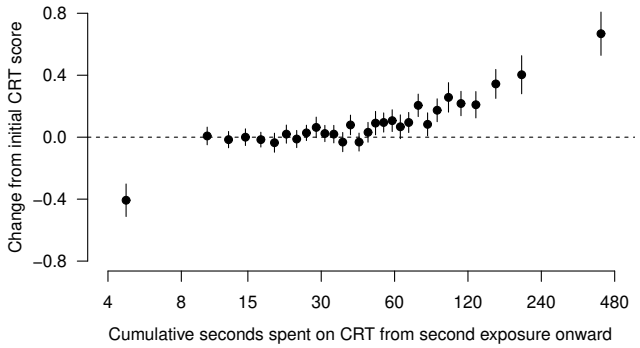

FIGURE 1: Time spent on CRT and score improvement. Analysis is of returning subjects within the Fall 2014 series. Data are sorted by cumulative time spent after first exposure and separated into 30 segments of 253 observations. The position of each dot corresponds to the average cumulative time spent and score increase for that segment. Error bars are 95% confidence intervals.

and ε_n are both small), or if S_1 and $R_{2:n}$ measure the same thing and ε_n is small. Both of these conditions appear to be met. First, scores are highly stable: subjects miss 90% of problems they missed on the preceding trial, and solve 95% of the problems they solved on the preceding trial (see Appendix C for further analysis). Second, score increases appear to indicate reflection, as they are more likely among people who solved other items (see appendix D), and are limited to those who continue to spend time on the test upon re-exposure (see Figure 1 and Appendix E).[3] Moreover, this subset is not just discovering and memorizing the correct responses; they appear to be learning how to solve these types of problems, as their improvements transfer to a modified CRT with different correct answers (contradicting Chandler et al., 2013, see Appendix F).

In any case, secular trends in the predictive validity of some instrument are easy to test for: one can simply check whether the correlation of interest changes or not. We can perform a few such tests with our data. First, in our Fall 2014 studies, we obtained self-reported SAT scores from 1,407 MTurkers who took the CRT at least twice.[4] Their final CRT scores predict SAT about as well as their initial scores, and the changes in score add significant incremental validity (See Table 4 and Appendix G).[5] Second, 327 subjects from the Fall of 2013 returned in the Spring of 2014 where they encountered the Linda problem (Tversky & Kahneman, 1983), six items from Raven's Advanced Progressive Matrices (Raven, 1941), and the CRT (again). Once again, performance on these other tests was predicted as well by final CRT scores as by initial scores (see Appendix H). Additionally, using self-reports as a proxy for prior CRT exposure, Bialek & Pennycook (2017) find no evidence that the test's predictive validity decreases across a large battery of covariates.

Those who fret about the test's continued validity assume, reasonably, that someone who scores a 0/3 the first time but a 3/3 the second time, was originally correctly classified (as unreflective) and now misclassified (as reflective) and erroneously lumped with those who got 3/3 the first time.[6] At first blush, this concern seems warranted: parroting answers one learns is not the same as generating those answers oneself. But suppose such a person had misgivings about their

[3]We emphasize that time spent on subsequent exposure predicts *improvement* in CRT score. The underlying relation between time spent on the CRT and CRT score is actually negative in these data.

[4]Of the 14,500 responses in this survey, 7,339 included SAT scores for both subject tests. In order to identify and omit spurious reports, respondents were not informed that scores range from 200 to 800, and we deleted 1,135 score reports that fell outside of that range. If individuals reported legitimate, but *different* SAT scores on different occasions, we averaged them. After this kind of cleaning, self-reported SAT scores typically correlate very highly with actual SAT scores (Kuncel, Crede & Thomas, 2005).

[5]Using modified CRT items that subjects had not seen before, Baron and co-authors (2014) report that both CRT score and CRT response time can be used to diagnose reflective tendencies. We worry that response times may be less robust to prior exposure than scores, because repeated exposure has negligible effect on scores, but has massive effects on response times. Even upon *first* exposure to the CRT in our data, response times already appear to add no incremental validity, beyond the scores themselves, for predicting performance on SAT, Raven's, or the Linda problem.

[6]Though useful as a thought experiment, this event is extremely rare (in our Fall 2014 series). Of the 2022 instances in which someone scored a 0 out of 3 and returned to take it again, only 48 got a perfect score the next time.

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

TABLE 4: Mean SAT scores sorted by initial and final CRT scores # of scores

Final CRT	Initial CRT 0	1	2	3	Overall
0	1104 297	1201 32	1230 13	1446 7	1124 349
1	1097 38	1209 141	1225 25	1470 1	1192 205
2	1231 17	1240 54	1256 187	1293 34	1256 292
3	1147 11	1269 29	1323 61	1302 460	1300 561
Overall	1111 363	1222 256	1266 286	1304 502	1231 1407

answers, the curiosity to act upon this doubt by Googling these items, the patience to sit through YouTube tutorials explaining their solutions, and the ability to remember these solutions when they encounter those items again. Those faculties sound conceptually close to what the test is intended to assess, and possibly even a purer measure than the sum of traits that enable correct solutions the first time, which include facility with algebra and with puzzles. Thus, we can find merit in the opposite interpretation: that this person was initially misclassified as unreflective, and is now being correctly classified as reflective.

Although we've focused on the CRT, this underlying logic applies to the shelf-life of any test. If current performance is a faithful proxy for initial performance or if change in performance measures the same thing as initial performance, the test's predictive validity won't be harmed by repeated exposure. Indeed, Appendix I shows that average performance on the Raven's and Linda items are about as stable as CRT scores. Just as a wine may become better, worse, or different as it acquires and loses various chemical aspects, the quality of a test may change depending upon the amounts of various traits a correct response betokens and the exact relations between levels of those traits and other constructs of interest (e.g., risk preferences, trolley preferences, authoritarianism, belief in God, and so on).[7]

The foregoing discussion should give pause to those who assume that the psychometric value of the CRT (or any test) necessarily declines with time. This could occur, but there is no compelling reason to think it is typical. Moreover, two primary concerns associated with the continued use of any test – response variance and predictive validity – can be straightforwardly assessed by simply looking at the data.[8] With respect to the CRT, that assessment will likely prove reassuring: in the most heavily exposed population, scores exhibit ample variance, are surprisingly stable, and retain their predictive validity, even when they change.

[7]Repetition of a test is just one of *many* factors that could affect performance. One could encourage people to take their time, warn them that the test is more difficult than it appears, tell them what the answers are *not* (see, e.g., Meyer & Frederick, 2018), and so on. Any of these other variables could also increase or reduce the test's predictive validity, depending on the population sampled and the other construct(s) of interest.

[8]However, one can still ask whether the CRT measures what it was originally intended to (the "organic" or innate disposition to stop and think). This cannot be answered solely by appealing to data of the usual sort, since the construct(s) measured by a psychological instrument *could* shift over time without affecting test scores. As a thought experiment, suppose that people used to rate the quality of their relationships by considering how close they felt with their parents, but now do so by considering the *quantity* of recent sexual experiences. This shift in the meaning of responses *could* occur with no changes in the responses themselves nor their covariation with other traits of interest (e.g., amount of drinking or frequency of suicidal thoughts).

References

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284.

Bialek, M., & Pennycook, G. (2017). The Cognitive Reflection Test is robust to multiple exposures. *Behavior Research Methods*, 1–7.

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2015). Cognitive Reflection Test: whom, how, when. (No. 68049). University Library of Munich, Germany.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making*, *5*(3), 182–191.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers. *Behavior Research Methods*, *46*(1), 112–130. http://dx.doi.org/10.3758/s13428-013-0365-7

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223–241.

Fernbach, P. M., Sloman, S. A., Louis, R. S., & Shube, J. N. (2012). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research*, *39*(5), 1115–1131.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*(4), 25–42.

Haigh, M. (2016). Has the Standard Cognitive Reflection Test Become a Victim of Its Own Success? *Advances in Cognitive Psychology, 12*(3), 145–149. http://dx.doi.org/10.5709/acp-0193-5

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373–385.

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

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judg-

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

ment. *Heuristics and biases: The psychology of intuitive judgment*, 49–81.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*, 63–82.

Lubinski, D., & Humphreys, L. G. (1990). A broadly based analysis of mathematical giftedness. *Intelligence*, *14*(3), 327–355.

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., ... & Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General, 144*(2), e16–e30. http://dx.doi.org/10.1037/xge0000049

Meyer, A., & Frederick, S. (2018). The bat and ball problem. *Unpublished manuscript*.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*(3), 335–346.

Pennycook, G., & Rand, D. G. (2018). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Available at SSRN: https://ssrn.com/abstract=3023545

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5*(3677), 1–12.

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *Psychology and Psychotherapy: Theory, Research and Practice, 19*(1), 137–150.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423–428.

Shenhav, A., Rand, D. G., & Data, J. D. G. (2017). The relationship between intertemporal choice and following the path of least resistance across choices, preferences, and beliefs. *Judgment and Decision making*, *12*(1), 1–18.

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision making*, *13*(3), 260–267.

Stieger, S., & Reips, U. D. (2016). A limitation of the Cognitive Reflection Test: familiarity. *PeerJ*, *4*, e2395.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making, 11*(1), 99–113.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293–315.

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta psychologica*, *134*(1), 16–28.

Appendix A: self-reported prior exposure probably reflects actual prior exposure plus latent ability and past performance

In the main text, we suggest that self-reported prior exposure to the CRT should not be interpreted as actual prior exposure or even as a noisy proxy for actual prior exposure. Here, we model it as a joint function of prior exposure, mental ability, and prior success on the test. First, we quantify the relation between likelihood of recalling prior exposure and amount of prior exposure. Then we attempt to differentiate two other determinants: mental ability and past performance on the test.

Table A1 shows how self-reported exposure varies according to how often subjects had encountered the CRT in the Fall 2014 series (reading down the columns) and how often they would (reading right along the rows). If self-reports accurately reflected the actual number of items respondents had seen before, it would increase to three by the second row and remain at three in all following rows. Though we do observe a large increase between the first and second rows, it does not go immediately to 3.0, but instead continues to rise gradually with further exposure. The increase moving right across the rows is most likely a composite effect of unobserved prior exposure and ability which facilitates memory.

Table A2 shows that people are more likely to recall their prior exposure to the test if they had done well on it ($r(6,761) = 0.25$, $p < 0.001$). This could either be interpreted as an effect of their prior success on their ability to recall the problems or as an effect of mental ability on both their prior success and their ability to recall the problems.

The first two columns of Table A3 show that the relation between previous performance and self-reported prior exposure is completely robust to controls for number of observed prior exposures and total number of exposures (as a proxy for unobserved prior exposure). The coefficient on previous CRT score barely changes with the addition of those controls.

The third column of Table A3 adds controls for subjects' previously reported number of items seen before to show that previous performance not only predicts cross-sectional differences in self-reported exposure, but also predicts changes in self-reported exposure within the same respondent.

Table A4 gives a more nuanced view of the average effect estimated in column 3 of Table 3. It shows the average self-reported number of items seen before, separately for each previous CRT score and previously reported number of items seen before.

A relation between mnemonic ability and general intelligence struggles to explain the fact that changes in previous performance continue to predict changes in self-reported prior exposure within the same subject over-time (even after controlling for number of prior exposures). This suggests

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

Table A1: # of CRT items reported seen before and $_{\text{\# of subjects responding}}$ across repeated testing.

nth appearance in series	# of times subject appeared in Fall 2014 series $_{\text{\# of subjects}}$						
	1 $_{3992}$	2 $_{1348}$	3 $_{614}$	4 $_{315}$	5 $_{183}$	6 $_{129}$	7+ $_{329}$
1st	1.36 $_{3653}$	1.62 $_{1244}$	1.73 $_{576}$	1.88 $_{290}$	1.84 $_{169}$	2.04 $_{116}$	2.25 $_{312}$
2nd		2.48 $_{1177}$	2.53 $_{521}$	2.59 $_{279}$	2.59 $_{160}$	2.55 $_{114}$	2.61 $_{298}$
3rd			2.68 $_{538}$	2.65 $_{277}$	2.70 $_{164}$	2.73 $_{113}$	2.67 $_{302}$
4th				2.76 $_{288}$	2.80 $_{162}$	2.81 $_{109}$	2.74 $_{287}$
5th					2.75 $_{159}$	2.74 $_{116}$	2.75 $_{300}$
6th						2.85 $_{116}$	2.78 $_{303}$
7th							2.81 $_{305}$

Table A2: relation between previous CRT score and self-reported prior exposure.

CRT score at time t-1 $_{\text{\# of observations}}$	mean # of CRT items reported seen before
0 $_{1667}$	2.36
1 $_{1049}$	2.52
2 $_{1404}$	2.76
3 $_{2643}$	2.87

some direct effect of prior performance on problem recall. But regardless of whether the relation is actually driven by past performance or merely by general intelligence, self-reported prior exposure will proxy for the ability to solve these problems above and beyond any effect of exposure, per se.

Appendix B: the relation between initial performance and frequency of later appearance

More frequent subjects in our study perform better on the CRT, even on their first exposure. To help differentiate selection effects from effects of unobserved prior exposure, we attempt to identify subjects who probably hadn't seen the CRT prior to our study by restricting analyses to those who (1) did not appear in any prior series of our data, (2) reported having seen zero items on their first exposure, and (3) reported having seen three items on every subsequent exposure (which provides evidence that their first report was accurate).

The positive relation between frequency of exposure and initial performance remains (p = 0.07) and is of similar magnitude to full sample estimates, suggesting that willingness to repeatedly engage in this task indicates greater aptitude for it, even if prior MTurk activity had not brought them in contact with the CRT. The more active subjects in our study were markedly less likely to be encountering the CRT for the first time in this study, suggesting a significant role of unobserved – and heavy – prior exposure.[9]

In the demographics section of the survey, subjects reported their SAT scores and educational attainment. Those who appear more frequently in our survey were more likely to report a valid SAT score (r(6,908) = 0.04, p = 0.002), and more likely to report having completed college (r(6,759) = 0.04, p = 0.001). However, there was no significant relation between frequency of appearance and the SAT score (r(2,920) = -0.00, p = 0.80).

Table B2 shows that the effects of repeated exposure on performance are similar across items (moving left to right within a row). The relation between frequency of appearance (moving down within a column) is strongest for bat and ball, followed by widgets, and weakest for lily pads (all three pairwise comparisons, p < .01).

Appendix C: response stability

Table 3 shows that average CRT scores don't increase much over time, but that could either indicate stability of response, or offsetting response variance (people who get it right forgetting and people who got it wrong improving, with similar magnitudes). Table C1 differentiates these possibilities by showing the probability of switching from wrong to right, and from right to wrong, at every possible transition. These probabilities are uniformly low which helps explain why the CRT maintains its predictive validity.

Table C2 differentiates the common or "intuitive" errors of 10, 100, and 24, from other "idiosyncratic" errors.

[9] You can also use this table to compare the effect of repeat exposure on CRT "virgins" to the effect of repeat exposure on others. CRT virgins improve by 0.066 items correct per exposure. Others improve by 0.023 items per exposure.

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

Table A3: OLS estimates of the effect of prior exposure and previous performance on self-reported number of items seen before $_{\text{standard error}}$.

	Dependent variable = # of items seen before		
previous CRT Score	0.17 $_{0.01}$	0.16 $_{0.01}$	0.06 $_{0.01}$
Constant	2.39 $_{0.03}$	.	.
Non parametric control for # of prior appearances during Fall 2014	No	Yes	Yes
Non parametric control for # of total appearances during Fall 2014	No	Yes	Yes
Non parametric control for previously reported # of items seen before	No	No	Yes
R^2	0.062	0.078	0.321
N	6,763	6,763	6,763

Table A4: relation between previous CRT performance and self-reported prior exposure, separately for each level of previous self-reported prior exposure.

CRT score at time t−1	# of CRT items reported seen before at time t−1 $_{\text{\# of observations}}$			
	0	1	2	3
0	1.51 $_{446}$	1.98 $_{100}$	2.55 $_{78}$	2.85 $_{837}$
1	1.69 $_{228}$	1.89 $_{73}$	2.55 $_{74}$	2.92 $_{608}$
2	2.10 $_{164}$	2.31 $_{88}$	2.48 $_{106}$	2.94 $_{979}$
3	1.99 $_{155}$	2.29 $_{96}$	2.63 $_{94}$	2.97 $_{2188}$

Although those who make intuitive errors (10, 100, 24) sometimes transition to idiosyncratic errors (e.g., 105, 20, 36), and those who make idiosyncratic errors sometimes transition to the correct answers (e.g., 5, 5, 36), idiosyncratic errors do not appear to function as a gateway to the truth. Of the 265 triplets with an idiosyncratic error in the middle position and a correct answer at the end, just 8% showed the pattern {intuitive→idiosyncratic→correct}, compared with 62% who merely "rediscovered" the truth {correct→idiosyncratic→correct}. Table C3 reproduces the analysis presented above at the item level.

Appendix D: people who initially solve more other items are more likely to improve

The main text asserts that more reflective individuals are more likely to improve CRT performance with further exposure. For each CRT problem, table D1 selects participants who initially got that problem wrong, separates them by their initial performance on other CRT problems and shows their rate of improvement with further exposure. In all cases, those who initially get more other items correct are more likely to improve.

For each CRT problem, table D2 selects subjects who initially got that problem right, separates them by their initial performance on other CRT problems and shows their rate of decrement with further exposure. In all cases, those who do better on other problems initially are less likely to get worse.

Table D3 makes linear assumptions on the rate of improvement and the change in rate of improvement to estimate the overall relation between rate of improvement and initial performance on other items for each of the three items. For all three items, people who initially get a given problem wrong are more likely to get it right later if they initially got other problems right.

Table D4 performs the same analysis among those who initially got each item right. It shows mixed results. For two out of the three items, better initial performance on other problems predicts a better chance of continuing to get the target problem correct. For the third problem, this relation reverses, but does not attain statistical significance.

Appendix E: people who continue to spend time are more likely to improve

The main text reports a strong relation between score improvement and the log of time spent on subsequent exposure ($r(7,487) = 0.21$). It also mentions that this does not reflect an underlying positive relation between time spent on the CRT and performance. In fact, that relation is negative, both overall ($r(14,272) = -0.14$), and excluding first observations ($r(7,433) = -0.12$). Further, the relation between score improvement and time spent on subsequent exposures is robust to controls for initial time spent (partial $r(7,450) = 0.19$).

We can distinguish two models of improvement in CRT score with repeat exposure: 1) between exposures, respondents encounter the answers in their daily lives, and 2) during each exposure, respondents think about the problems a little

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

Table B1: Mean CRT score among probable CRT "virgins" and $_{\text{mean CRT score of everybody else}}$.

# of times respondent appeared in our study	% encountering CRT first here	n^{th} appearance in our study						
		1st	2nd	3rd	4th	5th	6th	7th
1 (n=1535 $_{n=2457}$)	38%	1.04 $_{1.64}$						
2 (n= 206 $_{n=1142}$)	15%	1.17 $_{1.49}$	1.28 $_{1.56}$					
3 (n= 69 $_{n=545}$)	11%	1.32 $_{1.50}$	1.49 $_{1.53}$	1.52 $_{1.62}$				
4 (n= 33 $_{n=282}$)	10%	1.27 $_{1.64}$	1.36 $_{1.70}$	1.67 $_{1.71}$	1.61 $_{1.77}$			
5 (n= 12 $_{n=171}$)	7%	1.00 $_{1.65}$	1.00 $_{1.68}$	1.00 $_{1.75}$	0.92 $_{1.81}$	1.00 $_{1.81}$		
6 (n= 7 $_{n=122}$)	5%	1.14 $_{1.80}$	1.29 $_{1.81}$	1.29 $_{1.80}$	1.29 $_{1.85}$	1.29 $_{1.92}$	1.29 $_{1.99}$	
7+ (n= 14 $_{n=315}$)	4%	1.14 $_{1.78}$	1.50 $_{1.81}$	1.50 $_{1.81}$	1.64 $_{1.82}$	1.64 $_{1.85}$	1.71 $_{1.90}$	1.71 $_{1.91}$

Table B2: individual item solution rates across repeated testing.

# of times respondent appeared in our study:	% of subjects answering correctly on nth appearance in our study						
	1st	2nd	3rd	4th	5th	6th	7th
Bat and Ball							
1 $_{n=3992}$	40%						
2 $_{n=1348}$	41%	43%					
3 $_{n=614}$	43%	45%	46%				
4 $_{n=315}$	47%	47%	48%	50%			
5 $_{n=183}$	46%	48%	45%	46%	48%		
6 $_{n=129}$	59%	57%	57%	58%	59%	64%	
7+ $_{n=329}$	57%	57%	55%	56%	57%	59%	59%
Widgets							
1 $_{n=3992}$	45%						
2 $_{n=1348}$	46%	50%					
3 $_{n=614}$	49%	51%	55%				
4 $_{n=315}$	51%	55%	58%	59%			
5 $_{n=183}$	52%	54%	58%	59%	60%		
6 $_{n=129}$	59%	60%	58%	61%	63%	64%	
7+ $_{n=329}$	57%	59%	60%	60%	61%	62%	63%
Lily Pads							
1 $_{n=3992}$	56%						
2 $_{n=1348}$	57%	59%					
3 $_{n=614}$	55%	57%	61%				
4 $_{n=315}$	62%	64%	64%	66%			
5 $_{n=183}$	63%	63%	67%	70%	69%		
6 $_{n=129}$	59%	61%	61%	63%	67%	67%	
7+ $_{n=329}$	62%	64%	65%	66%	66%	68%	68%

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

Table C1: % probability of transitioning from wrong to right and from right to wrong.

# of times respondent appeared in our study	Transition between appearances in our study					
	1st to 2nd	2nd to 3rd	3rd to 4th	4th h 5th	5th to 6th	6th to 7th
2 $_{n=1348}$	13 $_{9}$					
3 $_{n=614}$	11 $_{8}$	13 $_{7}$				
4 $_{n=315}$	13 $_{7}$	9 $_{5}$	8 $_{3}$			
5 $_{n=183}$	12 $_{8}$	12 $_{6}$	9 $_{4}$	9 $_{6}$		
6 $_{n=129}$	8 $_{5}$	6 $_{5}$	9 $_{3}$	9 $_{2}$	7 $_{0}$	
7+ $_{n=329}$	9 $_{4}$	6 $_{4}$	6 $_{3}$	5 $_{2}$	7 $_{1}$	6 $_{3}$

Table C2: Percentage giving each type of answer on the next trial, conditional on type of answer given on the current trial.

	Answer on trial n+1		
Answer on trial n	Intuitive	Idiosyncratic	Correct
Intuitive $_{n=7736}$	87%	6%	7%
Idiosyncratic $_{n=2243}$	16%	64%	20%
Correct $_{n=12791}$	1%	3%	95%

Table C3: Percentage giving each type of answer on the next trial, conditioned on type of answer given on the current trial.

Answer on trial n:	Answer on trial n+1		
	Intuitive	Idiosyncratic	Correct
Bat and Ball			
Intuitive $_{n=3146}$	89%	5%	6%
Idiosyncratic $_{n=599}$	18%	54%	28%
Correct $_{n=3845}$	2%	6%	92%
Widgets			
Intuitive $_{n=2553}$	86%	6%	8%
Idiosyncratic $_{n=834}$	15%	67%	18%
Correct $_{n=4203}$	1%	3%	95%
Lily Pad			
Intuitive $_{n=2037}$	86%	8%	6%
Idiosyncratic $_{n=810}$	14%	69%	17%
Correct $_{n=4743}$	1%	2%	97%

Table D1: % solving each CRT problem after missing it on 1st try *(among those appearing three or more times in Fall 2014 series)*

nth appearance in Fall 2014 series	Initial score on other CRT items # of people		
Bat and Ball	0 $_{448}$	1 $_{204}$	2 $_{156}$
1st	0%	0%	0%
2nd	6%	11%	21%
3rd	8%	15%	22%
Widgets	0 $_{448}$	1 $_{200}$	2 $_{102}$
1st	0%	0%	0%
2nd	6%	17%	23%
3rd	10%	26%	24%
Lily Pad	0 $_{448}$	1 $_{152}$	2 $_{39}$
1st	0%	0%	0%
2nd	8%	15%	23%
3rd	11%	22%	28%

more. One crude test to distinguish between these two models asks whether score improvements are best explained by total weeks elapsed between exposures or by total minutes elapsed during exposures.

Table E presents the results of this test: specifically the expected score improvement (current score minus initial score) with each doubling of each independent variable. The constant in column 1 shows that one minute of additional reflection is associated with a score increase of about 0.15 items, and that each doubling of that time adds an additional 0.10 items correct, so that we would expect a respondent's score to exceed his initial score by 0.25 items after 2 minutes of time spent on re-exposure, by 0.35 after 4 minutes etc.... Column 2 presents the relation with weeks spent between exposures. It shows that we should expect scores to increase by 0.13 items correct when re-exposed one week after initial exposure, but only by another 0.03 for each doubling of that time, so that two weeks since initial exposure predicts a 0.16 item score increase and 4 weeks predicts a score increase of 0.19 items. Finally, column 3 models score improvement by number of previous exposures, as we do in our primary analysis. It shows that we should expect scores to increase by 0.09 items on first re-exposure, but only by 0.01 additional items for each additional doubling of exposures, such that 2 additional exposures predicts scores to increase by 0.10 items, whereas 4 additional exposures predicts a score increase of just 0.11 items.

One simple way to compare these models is by the percentage of variation in score change that they can explain. R2 of the "minutes spent" model is more than ten times higher than R2 of the "weeks passed" model. And R2 of the weeks passed model is itself almost ten times higher than R2 of column 3's "pure exposure" model. Another way to compare these models is to hold each constant and ask whether orthogonal variation in the other explains significant variation in the criterion. Columns 4 through 6 show that the

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

Table D2: % continuing to solve each CRT problem after solving it on 1st try *(among those appearing three or more times in Fall 2014 series)*

nth appearance in Fall 2014 series	Initial score on other CRT items # of people		
Bat and Ball	0 74	1 141	2 547
1st	100%	100%	100%
2nd	81%	84%	92%
3rd	81%	75%	91%
Widgets	0 78	1 195	2 547
1st	100%	100%	100%
2nd	81%	92%	97%
3rd	81%	92%	98%
Lily Pad	0 126	1 258	2 547
1st	100%	100%	100%
2nd	87%	95%	98%
3rd	93%	95%	98%

coefficient relating score change to time spent remains stable when controlling for weeks passed, but that the coefficient on weeks passed falls and even flips sign when controlling for time spent.

Appendix F: transfer of learning to modified CRT

If score improvements betoken continued reflection, subjects who improve on the test might not only learn the answers to these items, but also acquire the concepts required to solve them. We test that prediction by examining how exposure to the standard CRT during the Fall of 2014 affects performance on a modified CRT (Table F, left most column) that 4,670 subjects encountered during the Winter of 2015. Initial scores on the modified CRT were higher among the 1,610 subjects who had previously been exposed to the standard CRT than among the 3,060 who hadn't (1.61 vs. 1.35, $p < 0.001$). Further, among the 1,028 subjects who were exposed to the standard CRT multiple times, improvement over the course of exposures predicts modified score over-and above initial score (partial $r = 0.44$, $p < 0.001$), and modified score is better predicted by final standard score than by initial standard score ($r(1,028) = 0.80$ vs. $r = 0.76$, $p < 0.001$). This confirms that the modest effects of repeat exposure go beyond a rote memorization of answers, and, in conjunction with the response time evidence, suggests that cognitive reflection may be captured as well by final score as by initial score. Table F presents item level results.

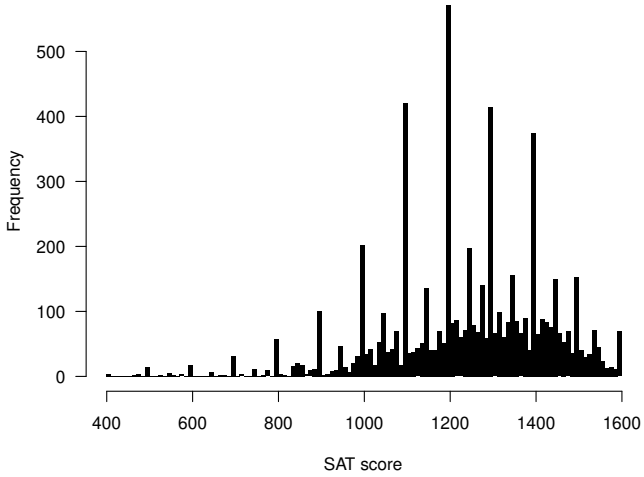

Figure G: Histogram of self-reported SAT scores.

Appendix G: SAT scores

Self-reported SAT score is the sum of self-reported quantitative and verbal sub-scores. The distribution is presented in Figure G. Verbal and quantitative sub-scores correlate strongly with each other ($r = 0.51$), and each is significantly related to CRT. Quantitative scores correlate somewhat more strongly ($r = 0.37$) than verbal scores ($r = 0.21$), but verbal scores are a significant predictor of CRT even after controlling for quantitative score.

The main text reports that SAT scores are just as well explained by Fall 2014 initial CRT scores as by Fall 2014 final CRT scores ($r(1,405) = 0.38$ vs. 0.36), and that final CRT adds incremental predictive validity over and above initial CRT score (partial $r(1,404) = 0.08$, $p = 0.002$).

Only 45% of those who appeared more than once in our study reported the same SAT score every time. While a few of the other 55% may have taken the SAT again in the interim and are reporting their latest score, for most, the variation reflects imperfect memory or insincere responding. In any case, our aforementioned finding that the relation between CRT and SAT is equally strong whether respondents are seeing the CRT for the first time or the nth time is essentially unchanged whether we just average the reported SAT scores (as we do above) or exclude the 55% who did not report the same score every time we asked them ($r(653) = 0.39$ vs. 0.36). However, although the partial correlation between final CRT and SAT after controlling for initial CRT does not change very much, it falls to insignificance in this smaller sample (partial $r(652) = 0.05$, $p = 0.203$). If we restrict our exclusions to respondents who report very different scores (a standard deviation greater than 100), we again find no significant decrease in the relation between SAT and CRT ($r(1084) = 0.37$ vs. 0.36), and we confirm the full-sample finding that final CRT score adds significant incremental validity over and above initial CRT score (partial $r(1,083) =$

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

Table D3: Probit estimates of the relation between initial performance on other items and rate of performance increase among those who initially got the target problem wrong $_{\text{standard error}}$.

	Target problem		
	Bat and Ball	Widgets	Lilypads
Number of other items initially solved * Number of prior exposures	$0.04_{0.01}$	$0.02_{0.01}$	$0.07_{0.02}$
Number of prior exposures	$0.14_{0.02}$	$0.16_{0.02}$	$0.16_{0.02}$
Number of other items initially solved	$0.29_{0.04}$	$0.34_{0.04}$	$0.25_{0.06}$
Non parametric control for # of total appearances during Fall 2014	Yes	Yes	Yes
N	7,762	7,236	5,939

Table D4: Probit estimates of the relation between initial performance on other items and rate of performance decrease among those who initially got the target problem right $_{\text{standard error}}$.

	Target problem		
	Bat and Ball	Widgets	Lilypads
Number of other items initially solved * Number of prior exposures	$-0.01_{0.01}$	$0.05_{0.02}$	$0.03_{0.02}$
Number of prior exposures	$-0.03_{0.04}$	$-0.22_{0.05}$	$-0.16_{0.05}$
Number of other items initially solved	$0.33_{0.06}$	$0.37_{0.07}$	$0.22_{0.08}$
Non parametric control for # of total appearances during Fall 2014	Yes	Yes	Yes
N	6,738	7,264	8,561

0.09, p = 0.002).

Table G1 takes a different approach. It estimates the correlation between CRT score and an individual's average reported SAT score, separately for each number of previous exposures within the study. A glance left-to-right within each row shows that there is no obvious decline in the CRT's predictive validity.

Table G2 formalizes this ocular analysis: it estimates the average change in the relation between mean reported SAT and CRT with each repeated exposure. Column 1 presents the univariate regression, which estimates an average SAT score of 1137 among 0s on the CRT and a 55 point increase for every additional CRT item solved. Column 2 adds non-parametric controls for number of times a subject appears in our data and the interaction between that control and CRT score. These controls are the equivalent of breaking the table into separate rows by total number of appearances in our data. They distinguish time-invariant covariates of frequent participation from effects of previous exposure. Column 3 adds number of previous exposures and the interaction between CRT score and the number of previous exposures. The interaction coefficient (0.3) estimates the average change in the relation between CRT and SAT with each additional exposure. It is small relative to the average relation (55), and statistically indistinguishable from 0. Further, comparing R2 between model 2 and model 3 shows that allowing the relation between CRT and SAT to vary with previous exposure did not improve model fit.

Even if the CRT continues to measure the same underlying trait, such that uniform prior exposure has no effect on its predictive validity, heterogeneous prior exposure could still be corrosive, as test scores alone would not differentiate between attaining a certain score on the first try and attaining that same score with the slight benefit of prior exposure. However, this effect is trivial. When we demean CRT scores by level of prior exposure, their ability to predict SAT scores barely increases (r=0.34 vs. 0.33).

Appendix H: Raven's and Linda

Our studies in Spring 2014 included two other cognitive tests: a six-item battery of Raven's Advanced Progressive Matrices (Raven, 1941), and Tversky and Kahneman's "Linda" problem (Tversky & Kahneman, 1983). Raven's advanced progressive matrices are a pattern matching task that is meant to assess fluid intelligence. The Linda problem presents subjects with a description of a woman who seems like a feminist, and asks the respondent whether she is more likely to be a feminist bank teller, or just a bank teller (whether or not she's a feminist). Many respondents commit the "conjunction fallacy" by choosing feminist bank teller over bank teller, and implying that the joint occurrence of two possibilities is more likely than one of the possibilities itself.

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

Table E: OLS estimates of change in CRT with doublings of three variables $_{\text{standard errors}}$. Dependent variable = current CRT score minus initial CRT score.

	1	2	3	4	5	6
Constant	0.15 $_{0.01}$	0.13 $_{0.01}$	0.09 $_{0.01}$	0.13 $_{0.01}$	0.27 $_{0.01}$	0.30 $_{0.01}$
log(minutes spent on CRT since first exposure)	0.10 $_{0.01}$			0.11 $_{0.01}$	0.15 $_{0.01}$	0.15 $_{0.01}$
log(weeks passed since first exposure to CRT)		0.03 $_{0.01}$		−0.02 $_{0.01}$		0.02 $_{0.01}$
log(# of exposures to CRT since first exposure)			0.01 $_{0.01}$		−0.09 $_{0.01}$	−0.11 $_{0.01}$
R2	0.0468	0.0032	0.0004	0.0485	0.0640	0.0649
N	7,489	7,489	7,489	7,489	7,489	7,489

Table F: Effects of exposure to standard CRT on initial modified CRT score in Winter of 2015.

	% correct on initial appearance in Winter 2015 series by prior participation in Fall 2014 series		Among Ss in Fall 2014 at least twice (N = 1028), correlation between initial Winter 2015 score and. . .	
Modified item text used in Winter 2015 series	Not in Fall (N = 3060)	In Fall 2014 (N = 1610)	Initial fall 2014 score	Final fall 2014 score
A bat and a ball cost $110 in total. The bat costs $98 more than the ball. How much does the ball cost? _____ dollars	33%	40%	0.53	0.55
If it takes 10 machines 10 minutes to make 10 widgets, how long would it take 50 machines to make 50 widgets? _____ minutes	51%	59%	0.65	0.68
In a lake there is a patch of lily pads. Every day, the patch doubles in size. If it takes 46 days for the patch to cover the lake, how long would it take for the patch to cover half the lake? _____ days	52%	63%	0.67	0.72

Table G1: Correlations between CRT and average reported SAT across repeated testing.

# of times subject appeared in our study	nth appearance in our study						
	1st	2nd	3rd	4th	5th	6th	7th
1 $_{n=1642}$	0.32						
2 $_{n=\ 643}$	0.42	0.40					
3 $_{n=\ 295}$	0.42	0.43	0.37				
4 $_{n=\ 141}$	0.37	0.32	0.34	0.35			
5 $_{n=\ 99}$	0.30	0.26	0.27	0.30	0.28		
6 $_{n=\ 58}$	0.26	0.24	0.27	0.25	0.31	0.31	
7+ $_{n=\ 171}$	0.29	0.31	0.29	0.33	0.32	0.34	0.27

The main text reports that final CRT score predicts Raven's and Linda as well as initial CRT score (Raven's: r(317) = 0.45 vs. 0.43; Linda: r(238) = 0.13 vs. 0.15). After controlling for initial score, the change in CRT is itself a significant predictor of Raven's score (partial r = 0.20, p < 0.01), but not of correct responses to the Linda problem (partial r = −0.01, p = 0.90).

We rely exclusively on the (relatively small) overlap between the Fall 2013 and Spring 2014 samples because respondents in the Spring of 2014 (when Linda and Raven's were administered) received feedback immediately after completing the CRT (i.e., that the answers were not 10, not 100, and not 24), creating a confound between the effect of that feedback, and the effect of any further exposure to the CRT. Table H1 ignores this confound, and examines the larger overlap between the Spring of 2014 and Fall of 2014 samples. It reports the relation with CRT score at four points in time: before feedback in the Spring, after feedback in the Spring, on first exposure in the Fall, and on final exposure in the Fall. It shows some evidence that repeated exposure with feedback reduces the CRT's ability to predict Linda, but no evidence that it reduces its ability to predict Raven's score.

Table H2 isolates the unique predictive contribution of each of the four CRT exposures. CRT scores appear to explain unique variation in Raven's score on every elicitation, but only the pre-feedback CRT score explains significant unique variation in Linda solution. See Meyer and Frederick (2018) for further discussion of the effect of invalidating the intuitive errors on the CRT's predictive validity.

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

Table G2: OLS estimates of the effect of previous exposure on the relation between CRT score and SAT score (dependent variable) standard error.

	Dependent variable = SAT Score		
CRT Score	55.1 $_{1.9}$	54.4 $_{3.9}$	54.4 $_{3.9}$
# of previous exposures	.	.	−1.7 $_{2.8}$
CRT Score × # of previous exposures	.	.	0.3 $_{1.2}$
Constant	1137.3 $_{4.0}$	1149.4 $_{7.8}$	1149.4 $_{7.8}$
Non-parametric control for total # of times respondent appeared in our study	No	Yes	Yes
Non-parametric control for total # of times respondent appeared in our study * CRT score	No	Yes	Yes
R2	0.1134	0.1345	0.1346
N	6,817	6,817	6,817

Table H1: correlations with CRT score at four different points among the subset of subjects who appeared in both the Spring and Fall 2014 studies t-statistic comparing correlation to spring 2014 pre-feedback correlation.

	Spring 2014 Pre-feedback CRT	Spring 2014 Post-feedback CRT	Fall 2014 initial CRT	Fall 2014 final CRT
Spring Raven's score (n = 1265)	0.38	0.39 $_{0.3}$	0.40 $_{0.8}$	0.40 $_{1.0}$
Spring Linda solution (n = 1003)	0.24	0.22 $_{1.5}$	0.20 $_{2.1}$	0.19 $_{2.2}$

Table H2: OLS estimates of the partial contribution of each CRT exposure after feedback standard error.

	DV=Raven's solution rate	DV=Linda solution rate
Spring Pre-feedback CRT	0.03 $_{0.01}$	0.08 $_{0.02}$
Spring Post-feedback CRT	0.02 $_{0.01}$	0.00 $_{0.03}$
Fall initial CRT	0.01 $_{0.02}$	−0.00 $_{0.03}$
Fall final CRT	0.04 $_{0.02}$	0.02 $_{0.03}$
Constant	0.26 $_{0.01}$	0.13 $_{0.03}$
R2	0.184	0.060
N	1265	1003

Appendix I: generalizability

The small effects of repeated exposure are not unique to the CRT, nor to the MTurk environment. The Spring 2014 series collected 6,843 responses from 5,191 unique MTurkers and found that the probability of avoiding the conjunction fallacy in the Linda problem increased by just 2.1% per exposure while Ravens scores increased by just 0.035 items per exposure (out of a possible score of 6). In the common terms of standard deviations per exposure, these two tests show similar repeat exposure effects to the CRT: 0.046 for Linda, 0.021 for Raven's and 0.020 for the CRT. Table I1 shows these longitudinal effects on average performance.

Table I1: % of Raven's matrices correct and % avoiding conjunction fallacy in Linda problem.

# of times respondent appeared in our study	nth appearance in our study		
	1st	2nd	3rd
1 (n=4032 $_{n=3837}$)	49 $_{28}$		
2 (n= 764 $_{n=\ 610}$)	46 $_{33}$	46 $_{36}$	
3+ (n= 286 $_{n=\ 183}$)	43 $_{27}$	43 $_{30}$	44 $_{32}$

Table I2: mean scores on modified CRT and geometric mean seconds to respond.

# of times respondent appeared in our study	nth appearance in our study		
	1st	2nd	3rd
1 $_{n=3535}$	1.42 $_{72}$		
2 $_{n=\ 854}$	1.47 $_{63}$	1.49 $_{38}$	
3+ $_{n=\ 281}$	1.57 $_{55}$	1.65 $_{39}$	1.79 $_{35}$

Table I2 replicates our primary analysis of change in average CRT score across repeated testing, but across administrations of the modified CRT during the Winter 2015 series (see Table F of appendix F for modified CRT materials). It replicates the small effects of repeat exposure that we find on the standard CRT in the Fall 2014 series.

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press

Table I3: Mean CRT scores among those previously told that the intuitive answers are wrong among everybody else.

# of times respondent appeared in Fall 2014	nth appearance in Fall 2014						
	1st	2nd	3rd	4th	5th	6th	7th
1 (n = 584 $_{n=3408}$)	1.99 $_{1.31}$						
2 (n = 258 $_{n=1090}$)	2.06 $_{1.29}$	2.12 $_{1.38}$					
3 (n = 149 $_{n=465}$)	2.03 $_{1.30}$	2.01 $_{1.37}$	2.11 $_{1.45}$				
4 (n = 82 $_{n=233}$)	2.04 $_{1.45}$	2.06 $_{1.52}$	2.10 $_{1.57}$	2.12 $_{1.62}$			
5 (n = 45 $_{n=138}$)	2.62 $_{1.28}$	2.51 $_{1.36}$	2.64 $_{1.40}$	2.64 $_{1.46}$	2.58 $_{1.49}$		
6 (n = 46 $_{n=83}$)	2.17 $_{1.54}$	2.17 $_{1.57}$	2.15 $_{1.55}$	2.13 $_{1.65}$	2.13 $_{1.75}$	2.15 $_{1.84}$	
7+ (n = 134 $_{n=195}$)	2.25 $_{1.42}$	2.19 $_{1.52}$	2.17 $_{1.54}$	2.22 $_{1.54}$	2.23 $_{1.57}$	2.22 $_{1.67}$	2.22 $_{1.69}$

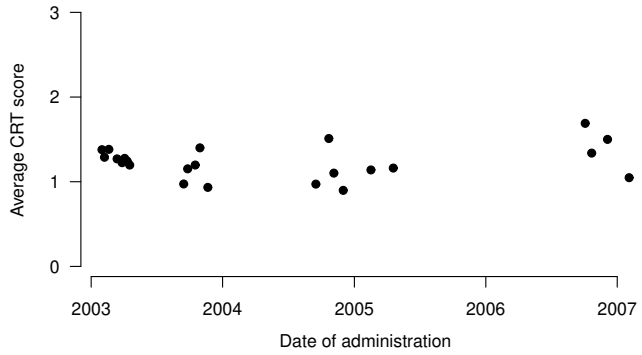

Figure I: University of Michigan CRT scores across repeated testing.

Further, the small effects of repeated exposure appear to generalize beyond MTurk. Although we didn't track individual subjects over time, we observed scores from 23 successive administrations of the CRT to a total of 1454 students on the University of Michigan campus, and see no evidence that aggregate scores improved there (see Figure I). Similarly, Brañas-Garza, Kujal, and Lenkei (2015) examine 118 administrations of the CRT, and, when they exclude MTurk studies, they find no statistically significant increase in solution rates from 2005 to 2014.

Although we have no reason to believe that the CRT is unique among cognitive tests or that MTurk is unique among experimental settings, important differences become apparent outside of experimental settings. In a meta-analysis of repeated exposure to tests used in organizational and educational settings, Hausknecht (2007) finds effects ten times larger (0.21 standard deviations per exposure).[10] The "discrepancy" between the tiny effects we observe and the modest effects observed elsewhere may be due partly to mean-reversion, as test takers in these other contexts are particularly likely to retake tests when they underperform expectations. Two more obvious reasons include higher performance incentives and explicit feedback after every exposure.

Although respondents don't typically get feedback about their CRT performance, 1298 people who participated in the Fall of 2014 surveys had previously participated in the Spring of 2014 surveys, which included a version of the CRT with partial feedback. Specifically, after those subjects responded to the CRT, they were told the most common errors on each problem (i.e., that the answers were not 10, not 100, and not 24), and received an opportunity to revise their responses. This feedback increased scores from 1.30 to 1.66 for those we never saw again and from 1.67 to 1.93 among those who returned for our fall study, where they averaged 2.07 items correct on first appearance. Thus, previous exposure with feedback (combined with the demand for an intervening response and a long delay) caused a 0.40 item increase, much larger than the 0.024 item average without feedback. Table I3 reproduces Table 3's analysis of previous exposure effects in the Fall of 2014 series, separately for those who had [and had not] previously participated in the Spring of 2014 study that provided feedback.

The 5612 respondents who hadn't appeared in the Spring 2014 series improved their CRT scores by about 0.037 items correct per exposure during the Fall 2014 series (standard error = 0.003), while the 1298 respondents who had appeared in the Spring of 2014 (which told them what the answers were not) only improved their CRT scores by about 0.007 items correct per exposure during the Fall 2014 series (standard error = 0.003). The data shown in Table 3 is the composite of these two groups.

[10]Note that this is *still* much smaller than the 0.72 standard deviations per exposure that self-reports of CRT familiarity imply.

https://doi.org/10.1017/S1930297500007683 Published online by Cambridge University Press