Reasoning as we read: how do readers understand conditional statements, implied meaning, and indirect meaning during comprehension?

Andrew Stewart

Division of Neuroscience and Experimental Psychology

@ajstewart_lang

https://github.com/ajstewartlang

Software Sustainability Institute

MANCHESTER 1824
The University of Manchester

It’s a matter of timing...

Much work in the area of human language processing focuses on *when* different kinds of information influence comprehension, and how readers build the mental representations associated with text meaning (situation models, e.g., Zwaan et al., 1995).

Reasoning as we Read

- Comprehension of conditionals is non-trivial.
- Amongst other things, it involves determining whether the conditional is describing a possible future situation, is a counterfactual requiring ‘undoing’ of the past. It involves determining the degree of belief of the conditional, deciding what speech act is being communicated (was that a promise?), and also the speaker’s persuasiveness when engaged in the use of a slippery-slope argument.

- Focus of today’s talk is on the research we’ve carried out examining how quickly all these things happen.
- My main interest on the moment-by-moment processing that accompanies the online comprehension of contextualised conditionals.
- Discussion as to what our results might mean for theories of how conditionals are produced and comprehended in everyday contexts.

Indicative vs. counterfactual conditionals

- Stewart, Haigh and Kidd (QJEP, 2009) showed that the interpretation of counterfactual conditionals but not indicative conditionals is constrained by prior context.
- Reading times to (e.g.) *If Darren had been athletic, he could probably have played on the rugby team* longer when prior context mismatches the presupposition.
- Indicative conditionals (e.g.) *If Darren is athletic, he probably plays on the rugby team* always quick to read.

Table 2. *Mean reading times and standard errors per condition for each of the three analysis regions in Experiment 2*

Conditional form	*Consistency*	*Region 1*	*Region 2 (critical region)*	*Region 3*
Counterfactual	Consistent	1,914 (82)	746 (39)	3,464 (156)
	Inconsistent	1,975 (81)	821 (43)	3,597 (149)
Indicative	Consistent	1,628 (71)	753 (41)	2,764 (123)
	Inconsistent	1,600 (69)	750 (38)	2,805 (165)

Note: Mean reading times in ms. Standard errors in parentheses.

Interaction in critical region (last word of antecedent, first word of consequent).

- All well and good but self-paced reading is a relatively coarse-grained measure of reading. Words appear one by one so normal reading processes are disrupted.
- Similar problem with examining event-related brain potentials during reading (e.g., Bonnefond & Van der Henst, 2013).
- So, if we want to measure how different sources of information influence the comprehension of conditionals during normal reading, how do we do it?

An eye-tracking in reading primer

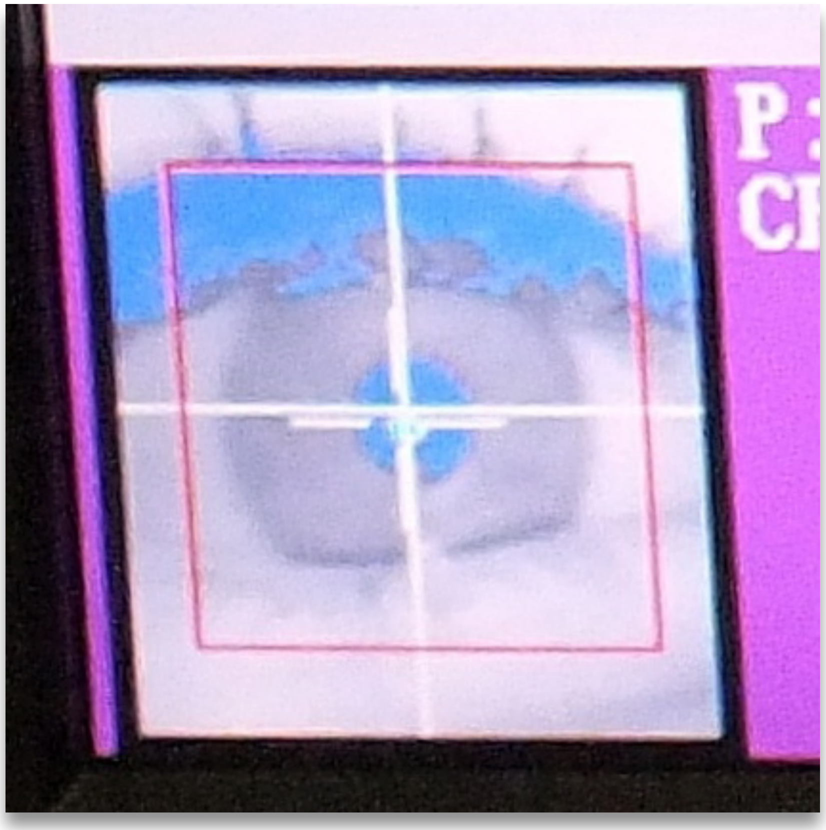
- Eye-movements during reading consist of fixations (for about 250 msec. each) and saccades (where the eye jumps from one location to another).

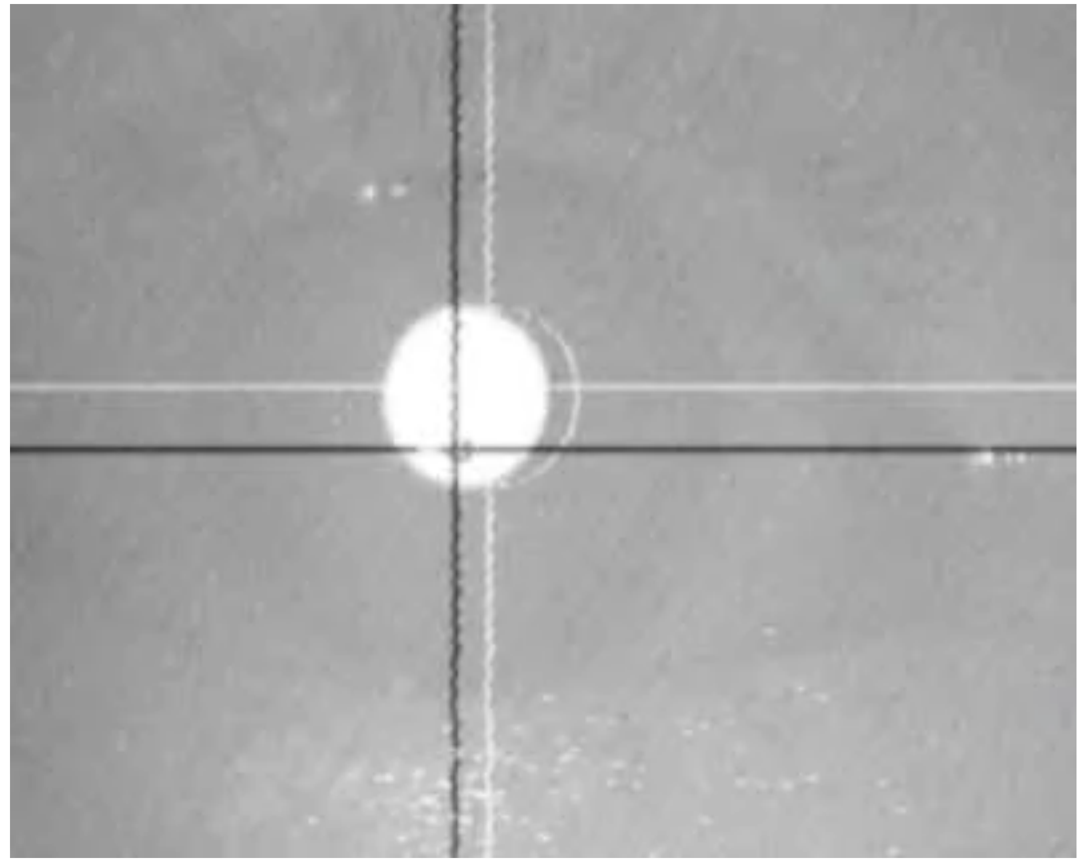
- During reading 10-15% of all eye-movements are backwards (called *regressions*) and they allow the reader to (re)look at previously read text.

- When fixating at a point in a word, you can actually see about 4 characters to the left and about 12-15 to the right of fixation. This is the *perceptual span* (McConkie & Rayner, 1975).

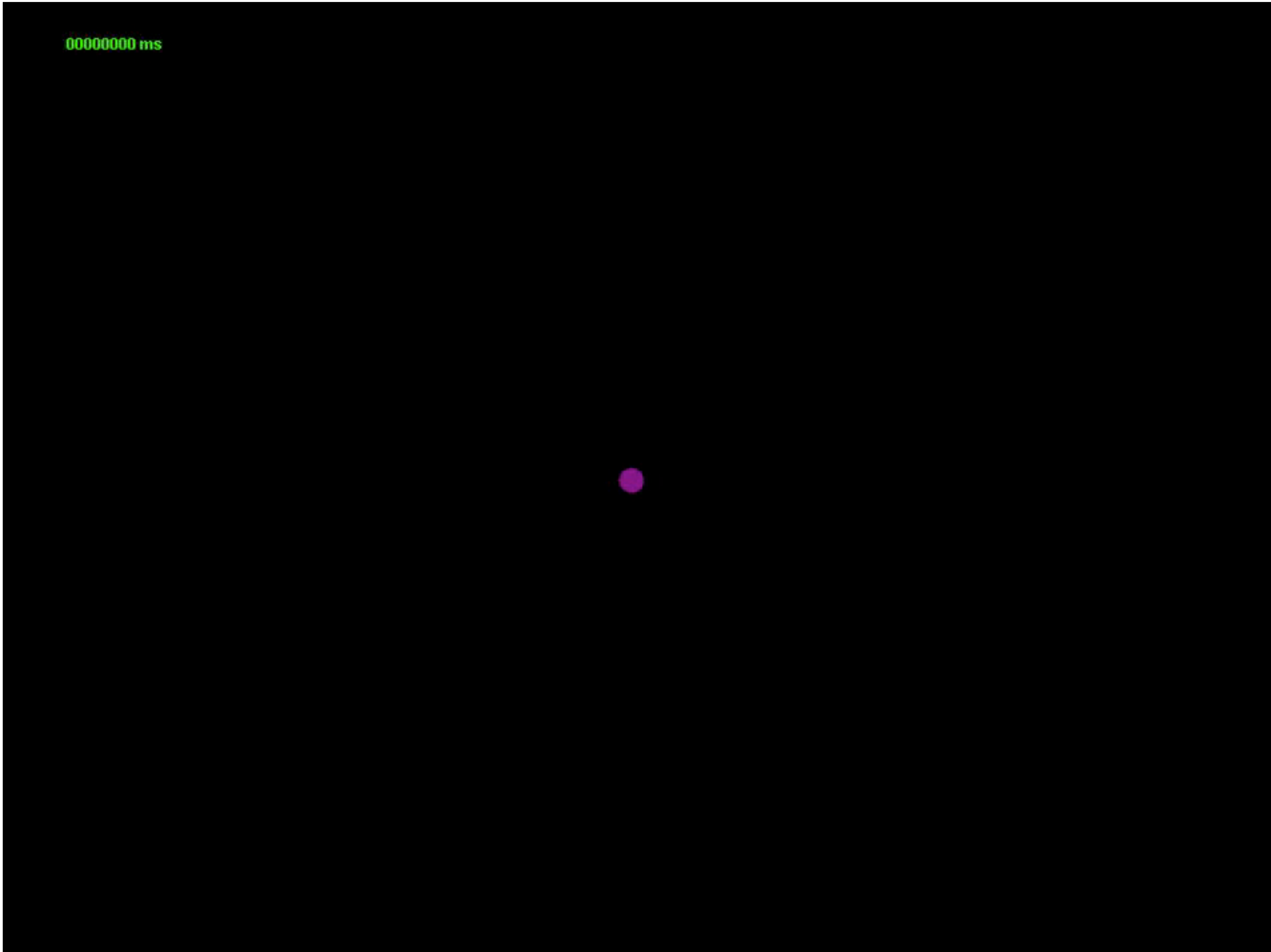
Infrared light is shone from the illuminator into the eye

Reflections of the infrared light from the eye are detected by the camera and overlaid on the image of the eye





Two reflections result - by measuring how these reflections move relative to each other, it's possible to calculate what the eye is looking at.


00000000 ms

Reasoning as we read

- **Reasoning** with conditionals as examined in the lab takes time (doesn’t always produce the ‘right’ answer) and takes effort.
- **Reading** of conditionals as examined in the lab is fast and (apparently) effortless.

If you want to lose weight, you need to exercise more.	**Advice**
If you wash my car, I'll pay you five pounds.	**Promise**
If you travel to Thailand, beware of pickpockets.	**Warning**
If you're late again, I'll fire you.	**Threat**

A framework for capturing indirect meaning…

Utility grid for a conditional promise:

A father saying to his son: If you wash my car, then I'll pay you a fiver.

{If	h	•	•	}
{Then	s	+	h	}
{	actor	utility	target	}

Utility grid for a conditional threat:

A traffic warden saying to motorist: If you park there, then I'll give you a ticket.

{If	h	•	•	}
{Then	s	-	h	}
{	actor	utility	target	}

Bonnefon (2009) Psych Review
Bonnefon, Haigh, & Stewart (2013) JML

But is this utility grid framework psychologically real?

Promises and threats differ in terms of the consequent event being positive or negative utility for the hearer.

Promises require the speaker to have control over whether the consequent event occurs (whereas tips do not).

How does a reader's knowledge of speaker control influence processing of the conditional?

If you submit your paper to the Journal of Physics, then I will publish it in the next issue.

Felicitous if uttered by someone who has control over what gets published, but infelicitous if uttered by someone who does not.

Stewart, Haigh, & Ferguson (2013) JEP:LMC

If you submit your paper to the Journal of Physics, then it stands a good chance of being published.

Felicitous regardless of the control the speaker has (as the consequent does not require the speaker to have control).

Does speaker control influence processing of the conditional as it is read?

Does this influence occur early or late?

Stewart, Haigh, & Ferguson (2013) JEP:LMC

We manipulated whether speaker did or did not have control of the consequent event and whether the conditional communicated a promise or a tip.

This gives us a 2 x 2 repeated measures design.

- Thirty six participants.
- Thirty two experimental vignettes.
- Thirty eight filler vignettes.
- Eye movements recorded using Eyelink 1000.

Stewart, Haigh, & Ferguson (2013) JEP:LMC

Alan had just presented his research paper to a meeting of leading physicists. During the coffee break he was called over by the Editor of the internationally renowned Journal of Physics/ by a junior colleague. The Editor/colleague was very impressed by Alan's findings and said that they should be widely publicised.

(a) As they parted, the Editor/colleague told Alan "if you submit your paper to the Journal of Physics, then I will publish it in the next issue".

(b) As they parted, the Editor/colleague told Alan "if you submit your paper to the Journal of Physics, then it stands a good chance of being published".

This comment made Alan consider his options carefully.

Stewart, Haigh, & Ferguson (2013) JEP:LMC

Analysis Regions

Pre-critical region	"if you submit your paper to the Journal of Physics,
Critical region	then I will publish it in the next issue".
Post-critical region	This comment made Alan consider his options carefully.

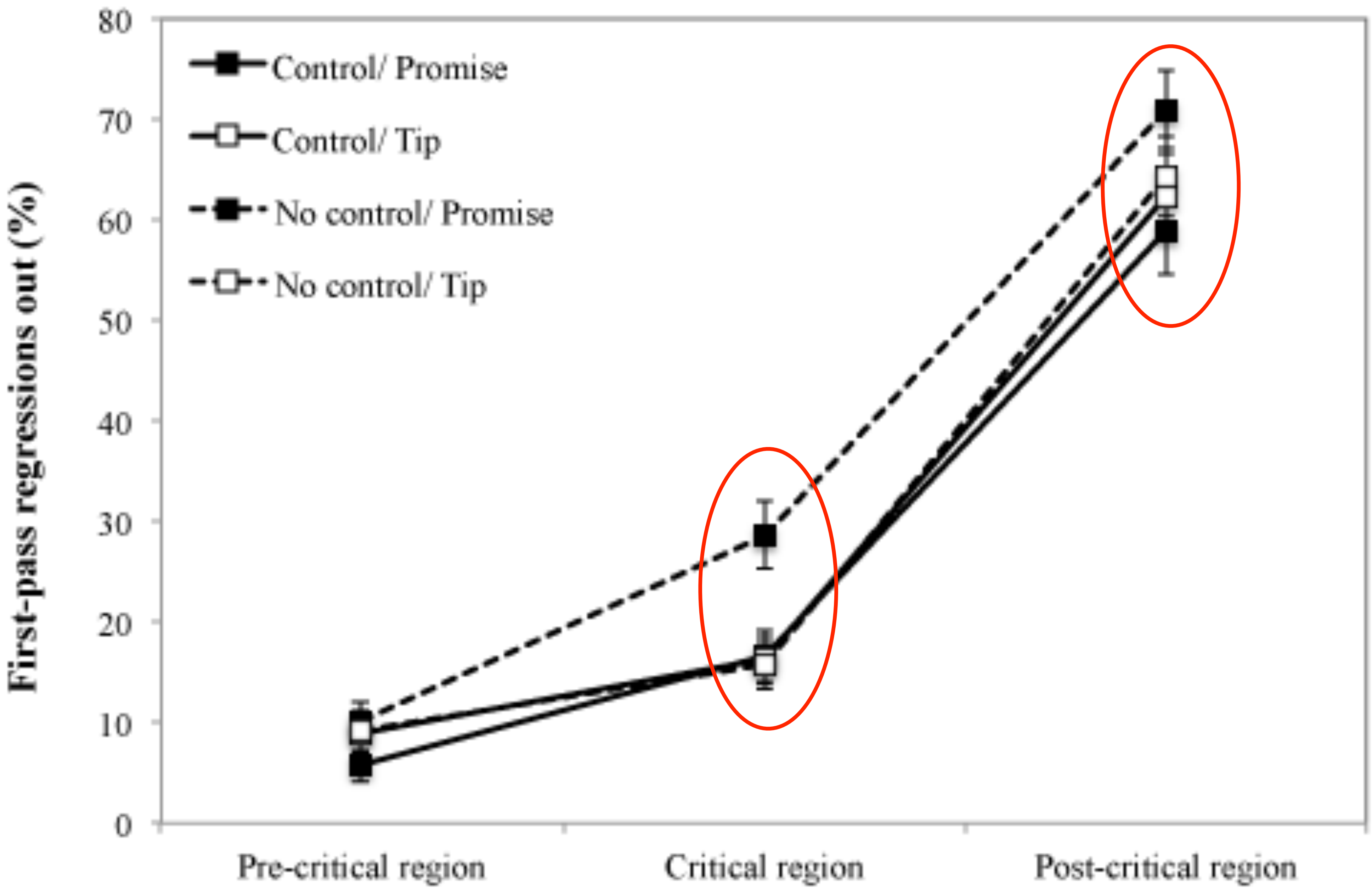
Stewart, Haigh, & Ferguson (2013) JEP:LMC

Measures

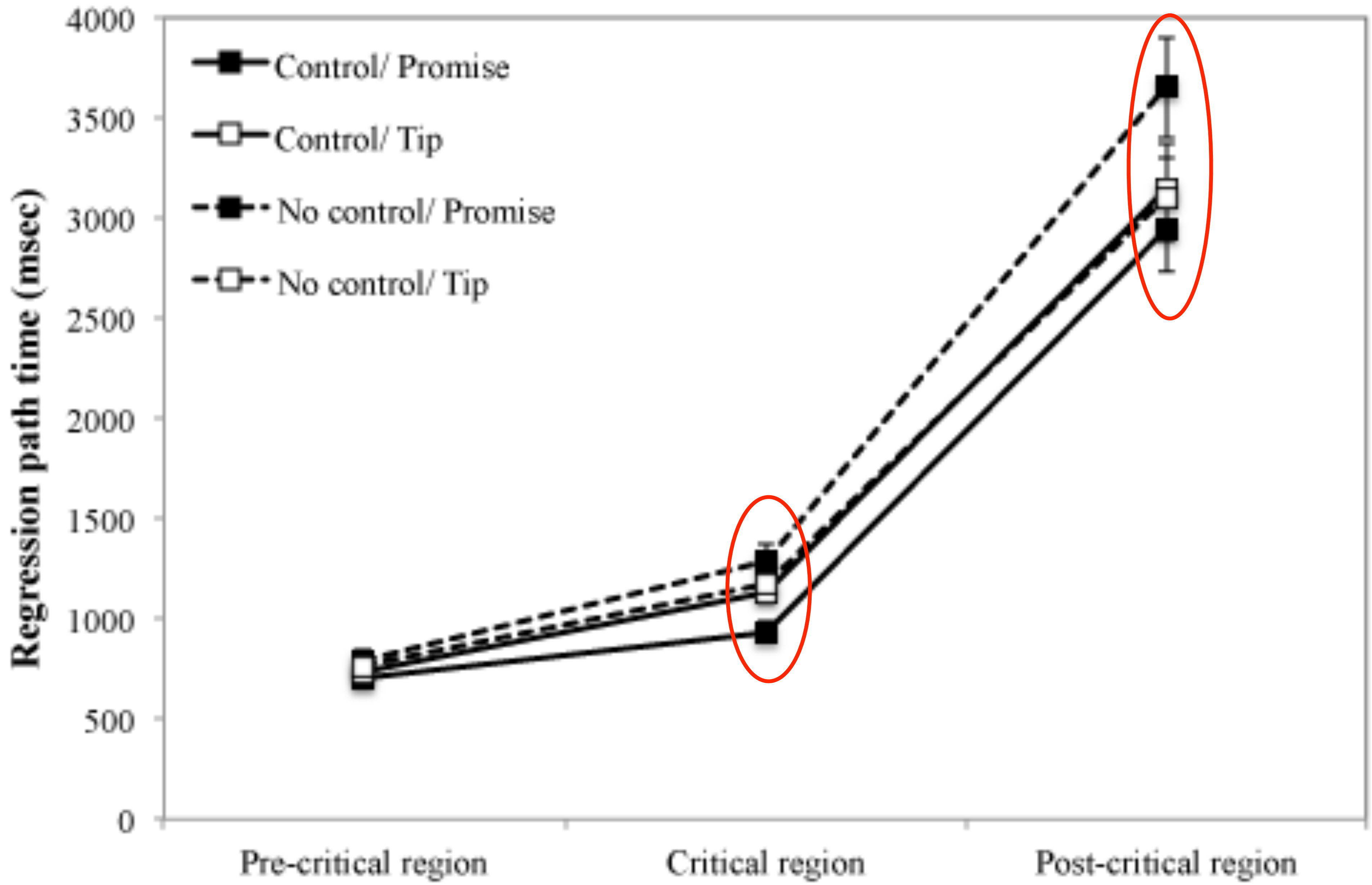
- First-pass regressions out (%): the degree to which left to right eye movements are disrupted while first reading a region of text.
- Regression path reading time (msec.): how long it takes a reader to go past a region of text after first entering it.
- Total reading time (msec.): sum of all fixation durations in a region.

Stewart, Haigh, & Ferguson (2013) JEP:LMC

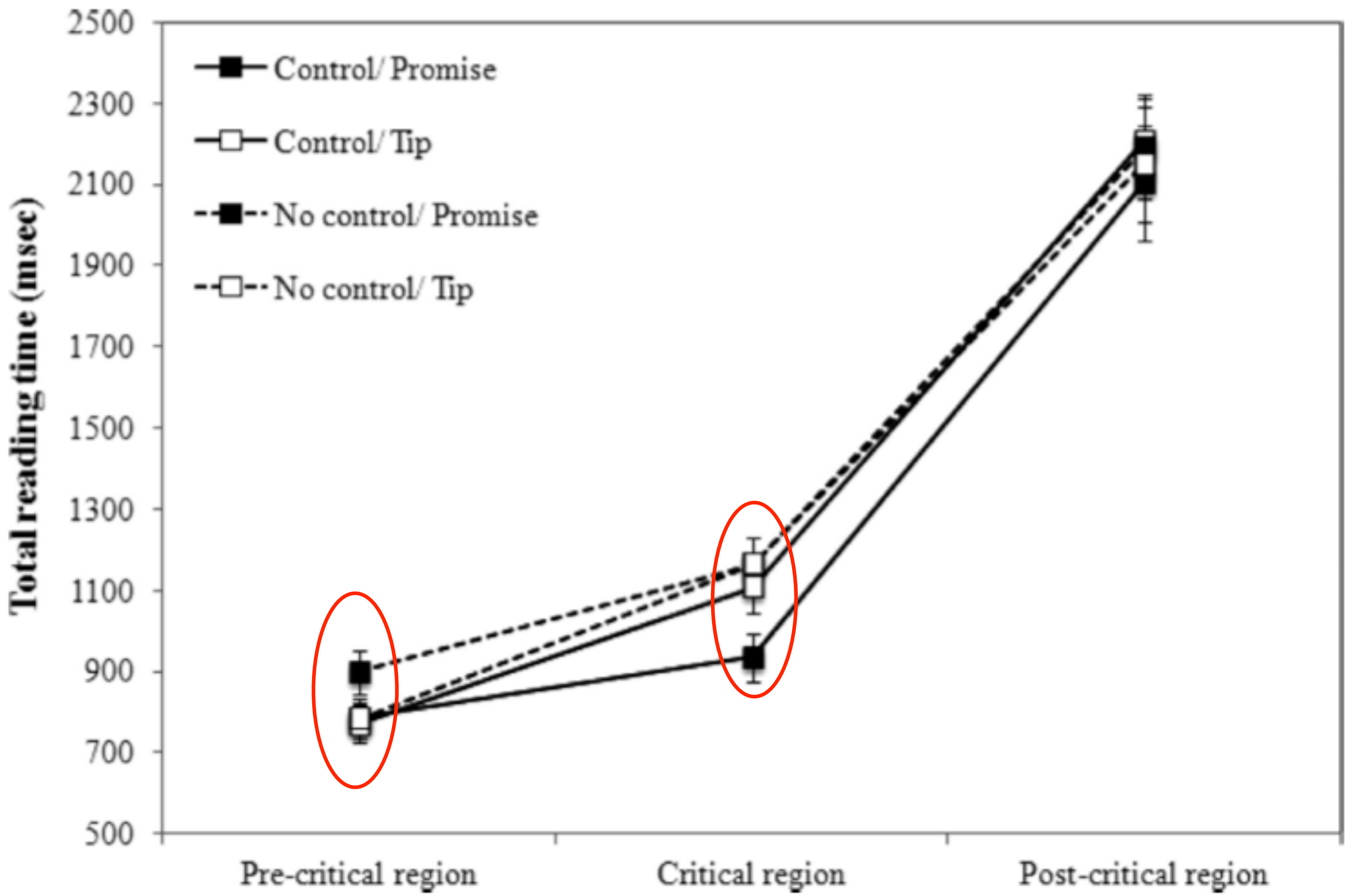
First Pass Regressions Out



Regression Path Time



Total Reading Time



- Promises are read slowly when the speaker isn’t in a position to make that promise. Tips are read at the same speed.

- Evidence on measures of processing that a reader’s knowledge of speaker control based on prior context and social cognitive information rapidly constrains the interpretation and representation of conditional events.

- Importantly, this effect emerges on measures associated with **early** processing.

- But is it as simple as this? Is conditional meaning simply mapped onto a utility grid?

- Consider Promises vs. Threat. Both are types of Inducement.

- Promises and Threats differ in their level of obligation - Promises are seen as having a greater obligation (Verbrugge et al., 2004, 2005).

- It seems possible to increase (but not decrease) the obligation of an Inducement.

- “If you do that again, I’ll put you in time out. That’s not a threat, that’s a promise….”

- Sounds ok.

- “If you do that again, I’ll put you in time out. That’s not a promise, that’s a threat….”

- Doesn’t sound ok.

- It seems that this allows threats to be subsumed under promises (i.e., they're a particular type of promise).

- But where's the evidence?

- Is it ok for a conditional threat to be referred to later in a text as "This promise", and is it NOT ok for a conditional promise to be referred to later as "This threat"?

We manipulated how promises and threats can be referred to later in a text.

This gives us a 2 x 2 repeated measures design.

- Forty participants.
- Thirty two experimental vignettes.
- Thirty two filler vignettes.
- Eye movements recorded using Eyelink 1000.

Wood, Haigh, and Stewart (2016) Exp Psych

Ian was at a builder's merchant to buy some paving slabs for a job. He approached the sales assistant intent on getting a good deal. She told him "if you buy in bulk, then I'll give you our trade discount" (Promise) / "if you only buy a small amount, then I'll stop your trade discount" (Threat). This promise / This threat helped Ian to make his decision. He thought about it for a while and then placed his order.

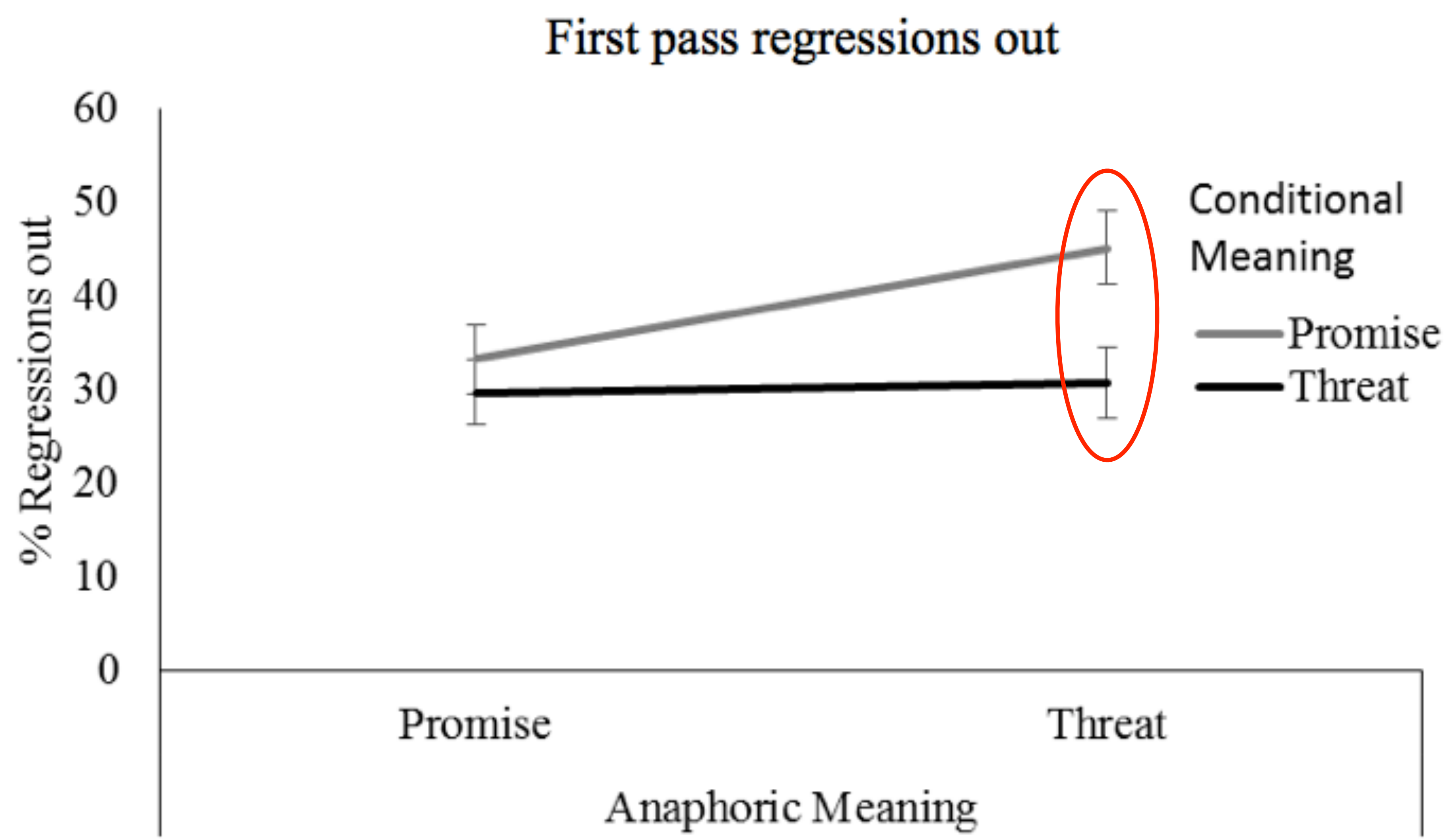
Critical region: This promise vs. This threat
Post-critical region: helped Ian to make his decision.

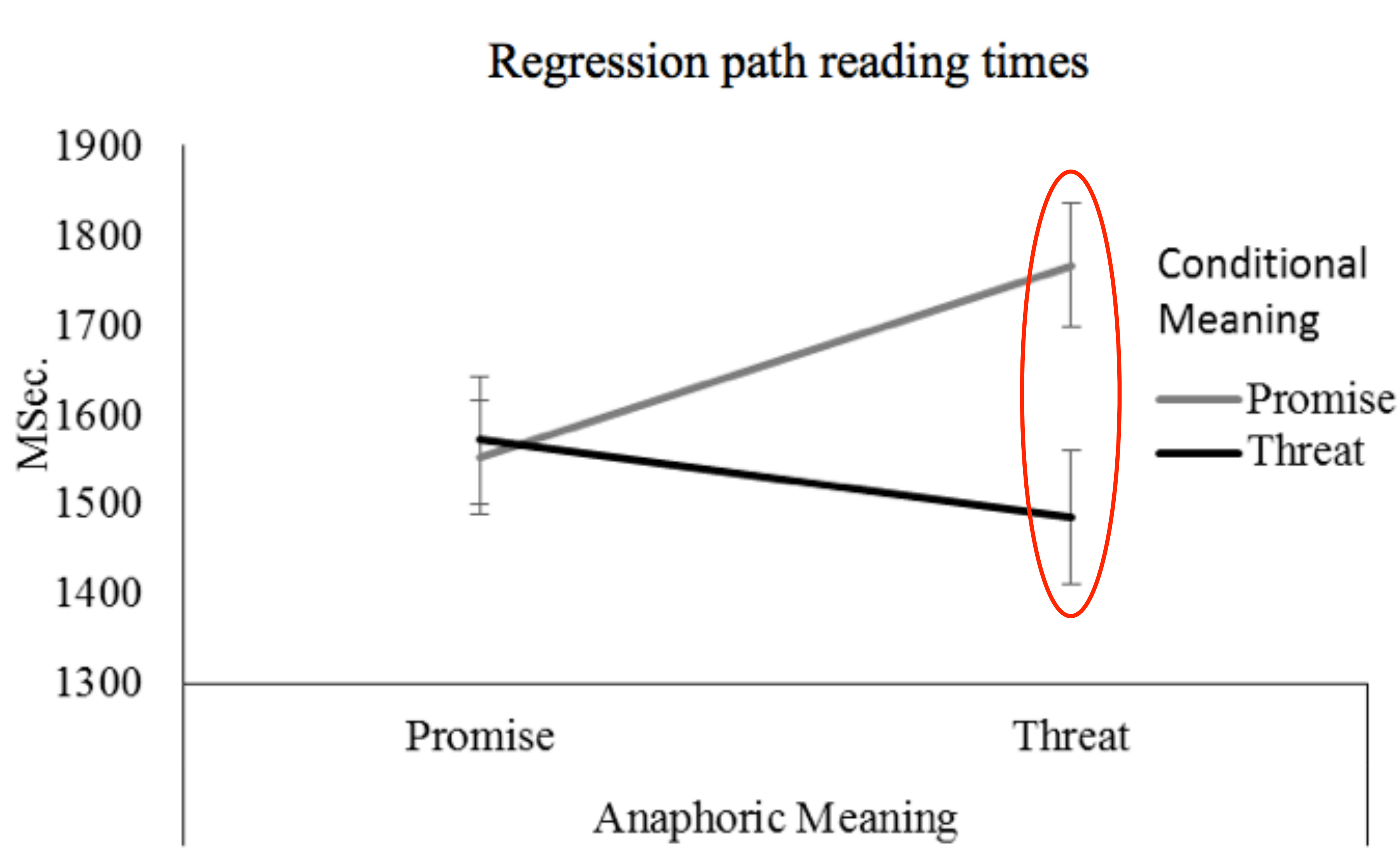
Wood, Haigh, and Stewart (2016) Exp Psych

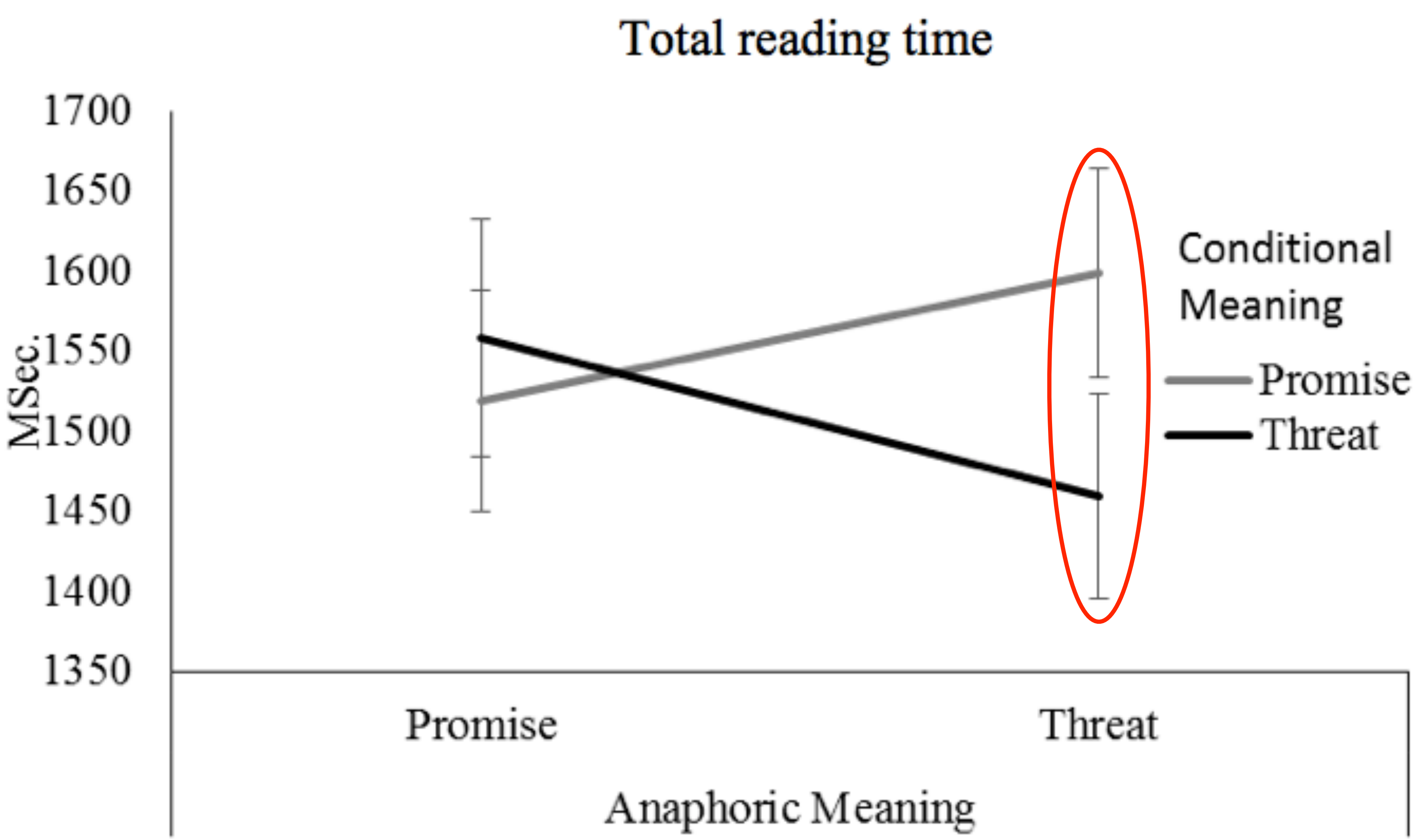
- No effects on Critical region, but clear effects on Post-critical region:

	Duration measures									Binomial measure		
	First Pass			Regression Path			Total Time			First Pass Regressions Out		
	b	SE	*t*	*b*	SE	*t*	*b*	SE	*t*	*b*	SE	*z*
Post-Critical Region												
Intercept	**981.4**	**67.27**	**14.59**	**1596.73**	**88.08**	**18.13**	**1536.34**	**94.07**	**16.33**	**-0.78**	**0.17**	**-4.54**
Conditional Meaning	-20.99	42.71	-0.49	**126.32**	**49.08**	**2.57**	46.15	42.06	1.10	**0.47**	**0.14**	**3.34**
Anaphoric Meaning	48.06	38.94	1.23	-59.41	42.87	-1.39	12.87	38.45	0.34	**-0.32**	**0.13**	**-2.47**
Interaction	-31.19	81.29	-0.38	**-308.14**	**84.41**	**-3.65**	**-187.47**	**86.25**	**-2.17**	**-0.53**	**0.26**	**-2.07**

Wood, Haigh, and Stewart (2016) Exp Psych


First pass regressions out
60
50
40
30
20
10
0
% Regressions out
Conditional
Meaning
Promise
Threat
Promise
Threat
Anaphoric Meaning


Regression path reading times
1900
1800
1700
1600
1500
1400
1300
MSec.
Promise
Threat
Anaphoric Meaning
Conditional Meaning
Promise
Threat



Total reading time
1700
1650
1600
1550
1500
1450
1400
1350
MSec.
Promise
Threat
Anaphoric Meaning
Conditional Meaning
Promise
Threat

- Data support the idea that readers are sensitive to pragmatic differences between Promises and Threats. Threats can be referred to as Promises, but Promises cannot be easily referred to as Threats.

- Conditionals are not used to just communicate implicit promise, tips, threats, and warnings.
- Conditionals are also used in *persuasion*.

Slippery Slope Arguments (SSAs)

If p, then q SSAs describe an initial proposal (P) and a predicted, undesirable consequence of this proposal (Q):

> *If voluntary euthanasia is ever legalised, then it will ultimately lead to the legalisation of involuntary euthanasia.*

Haigh, Wood, and Stewart (2016) Mem & Cog

SSAs can be thought of as a negative consequentialist argument (following Corner et al., 2011, Bonnefon & Hilton, 2004).

They work by implying something of the speaker's views, and inviting you to reject the initial proposal on the basis of what that might lead to.

If voluntary euthanasia is ever legalised, then it will ultimately lead to the legalisation of involuntary euthanasia.

In a paraphrasing study, we examined what SSAs are seen to reveal about the attitudes of the producer.

24 Ss presented with 24 SSAs and asked to write down what they think the producer believes.

Haigh, Wood, and Stewart (2016) Mem & Cog

Carly utters: *If voluntary euthanasia is ever legalised, then it will ultimately lead to the legalisation of involuntary euthanasia.*

~ 77% of responses indicate that participants inferred the speaker had a negative attitude towards the antecedent information.

Participant 2: “Carly disagrees with voluntary euthanasia”

Participant 3: “Carly does not think voluntary euthanasia should be legalised, as it could lead to murder.”

Participant 16: “Carly thinks the risks associated with the escalation of the laws is not worth legalising voluntary euthanasia.” [sic]

Participant 19: “Carly opposes voluntary euthanasia.”

Eye-tracking experiment

24 participants read 24 SSAs in one of three conditions: Speaker was known to be against the antecedent proposal (Consistent) *vs.* Speaker was known to support the antecedent proposal (Inconsistent) *vs.* Speaker's position towards the antecedent proposal was unclear (Neutral).

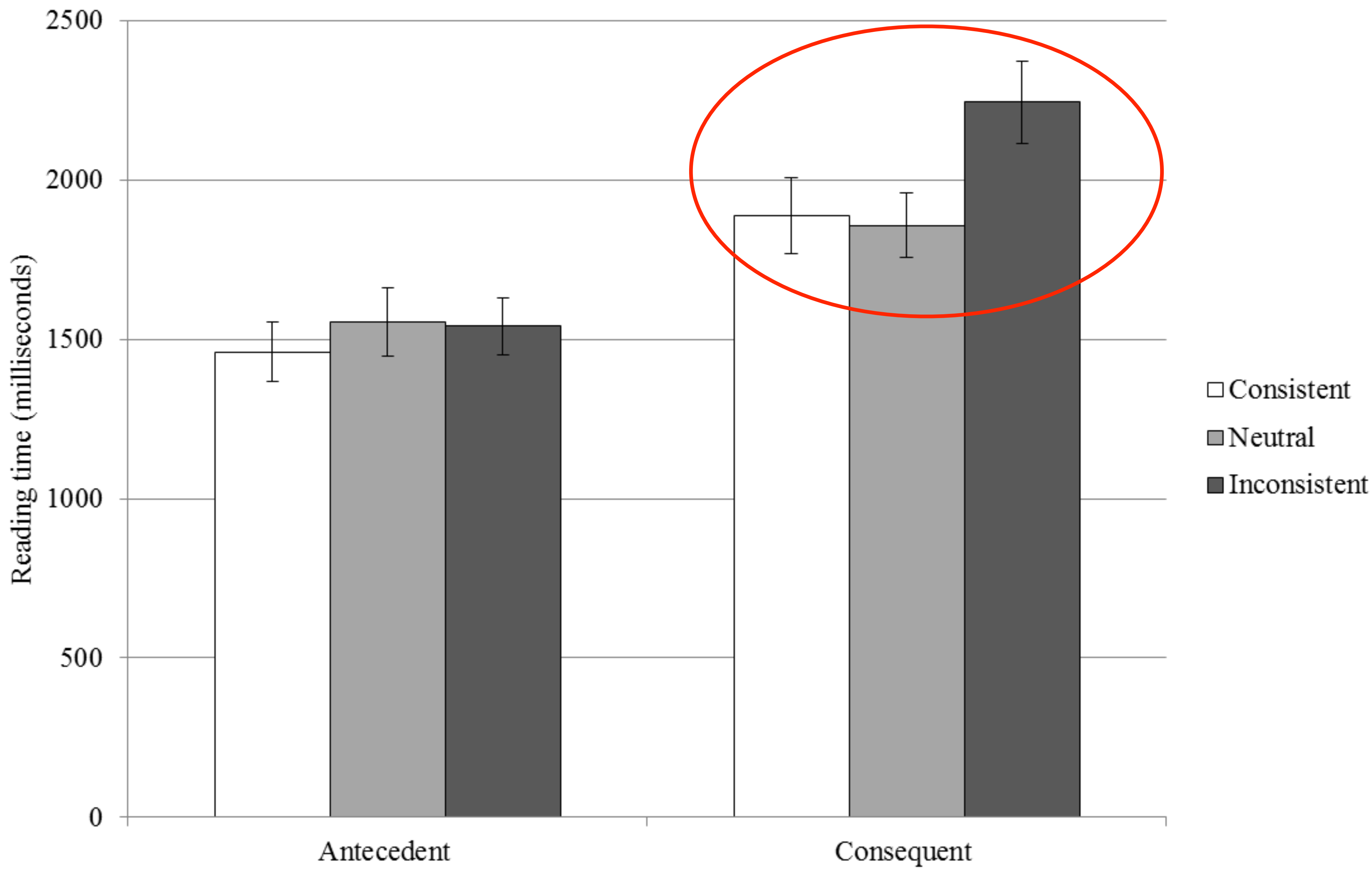
Haigh, Wood, and Stewart (2016) Mem & Cog

Introduction	Jayne and Carly were discussing their feelings about euthanasia.
a) Consistent context	*Carly was strongly against it becoming legal in the UK.*
b) Inconsistent context	*Carly was strongly in favour of it becoming legal in the UK.*
c) Neutral context	*Carly had recently heard that it could become legal in the UK.*
Antecedent (*A*)	She argued that |**"If voluntary euthanasia is ever legalised,|**
Consequent (*C*)	**|it will ultimately lead to the legalisation of involuntary euthanasia".|**
Final sentence	|They were both engrossed by a live television debate on the subject.|

Key analysis regions were the bold text in the Antecedent and Consequent.

Haigh, Wood, and Stewart (2016) Mem & Cog

Regression path reading times



Haigh, Wood, and Stewart (2016) Mem & Cog

The SSA eye tracking data are compatible with a view that readers have difficulty understanding a SSA when it goes against what it known about the producer's attitudes.

Readers are rapidly sensitive to the rhetorical function of SSAs and what they reveal about the producer's attitudes with respect to the antecedent proposition.

- Up to this point we have looked at meaning communicated indirectly by conditionals.

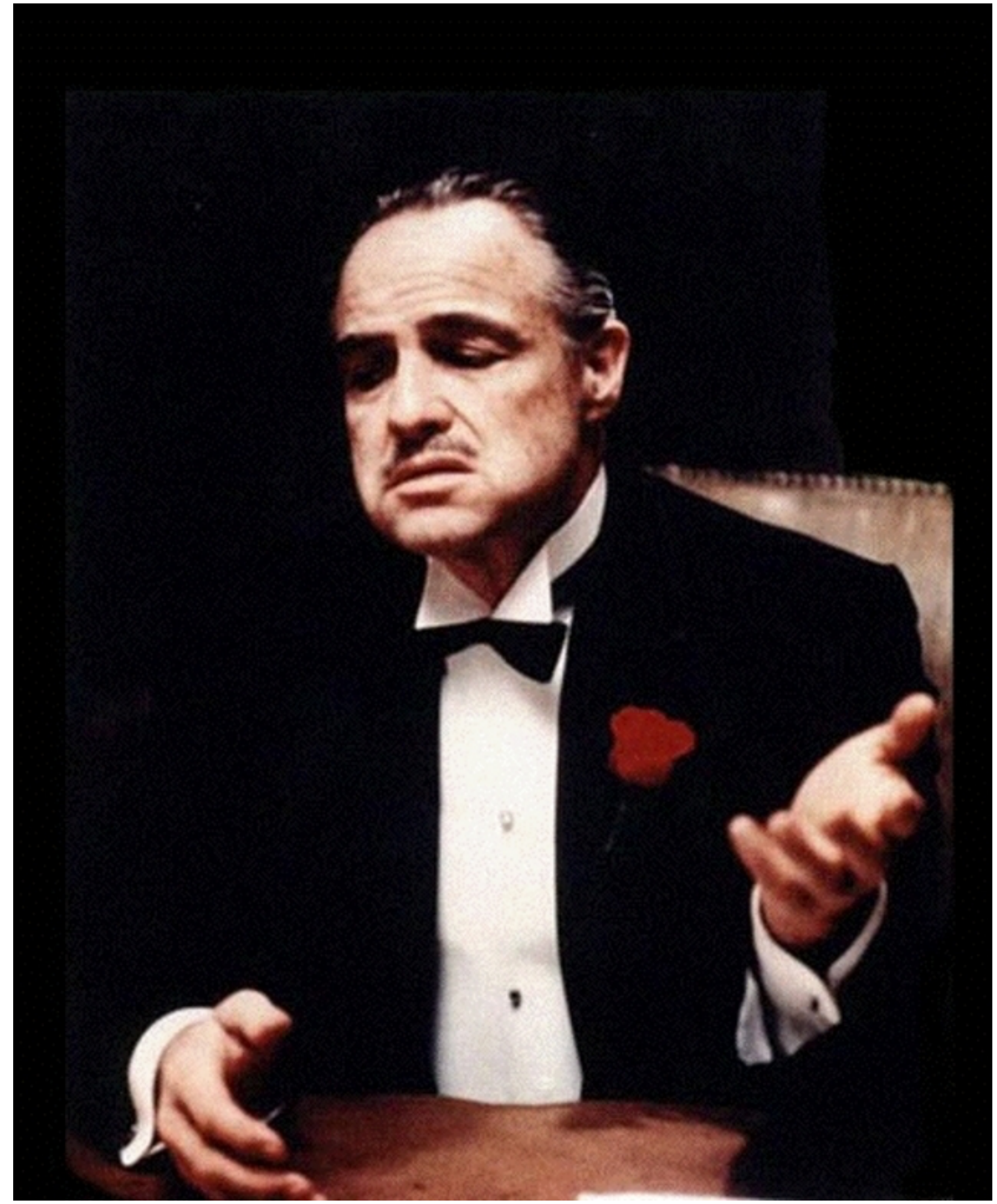
- Now indirect requests and replies…

Mitchell and Webb (2010)

Indirect Requests

- Long tradition in language research looking at conventionalised indirect requests such as *"Can you pass the salt?"*
- But many requests are non-conventionalised - they need context to be understood.
- Lee and Pinker (2010) developed the strategic speaker model to account for the use of indirect language. Key to this is *plausible deniability*.

Don Corleone: “I hear you’re the foreman of the jury in the Soprano trial. It’s an important civic responsibility. You have a wife and kids. We know you’ll do the right thing.”



Clearly an indirect threat/request that the listener finds the defendant not guilty. But with plausible deniability.

- A recipient’s face can be threatened if a request threatens their autonomy (i.e., has a high degree of imposition). If I know you’re going to agree to my request, I’m more likely to frame it politely (and indirectly).

Stewart et al. (2018) Discourse Processes

We manipulated the degree of imposition (high vs. low)
and the phrasing of the request (indirect vs. direct).

This gives us a 2 x 2 repeated measures design.

- Sixty participants.
- Twenty eight experimental vignettes.
- Fourteen filler vignettes.
- Eye movements recorded using Eyelink 1000.

Stewart et al. (2018) Discourse Processes

Doug was speeding in his car and was stopped by a traffic cop. Traffic cops in this area were known to be dishonest/honest. Doug said “Perhaps there is another way we can resolve this.”/”Doug said “I’ll give you £20 and you could let me go.” The cop accepted the bribe and Doug avoided the penalty. Doug was on his way to visit his grandmother.

Stewart et al. (2018) Discourse Processes

Three analysis regions:

Doug was speeding in his car and was stopped by a traffic cop. |Traffic cops in this area were known to be dishonest. IMPOSITION| Doug said |“Perhaps there is another way we can resolve this”. CRITICAL| The cop accepted the bribe and Doug avoided the penalty. POST-CRITICAL| Doug was on his way to visit his grandmother.

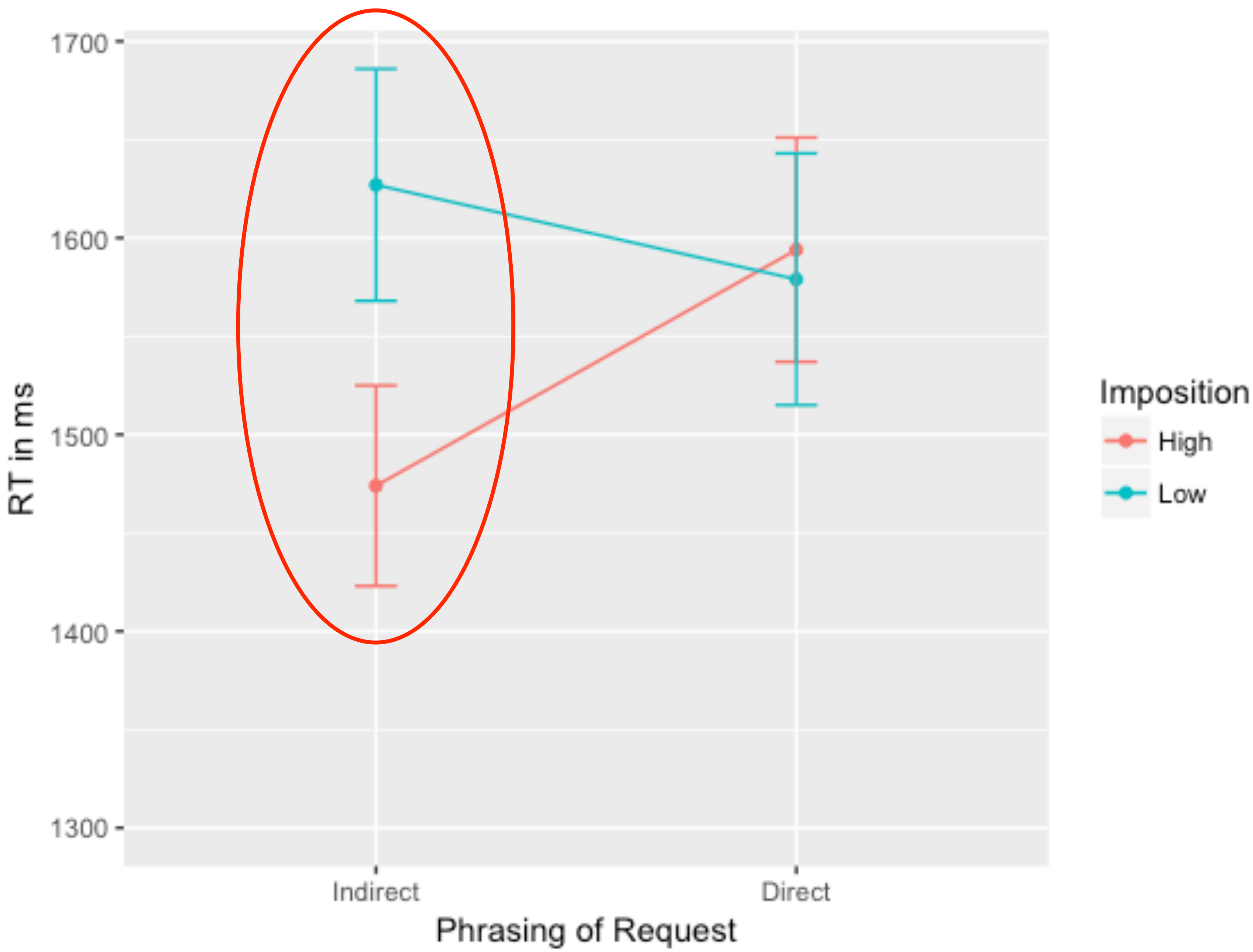
Stewart et al. (2018) Discourse Processes

- Clear effects on Critical region:

	Duration measures									Binomial measure					
	First Pass			Regression Path			Total Time			First Pass Regressions Out – by participants			First Pass Regressions Out – by items		
	b	SE	*t*	*b*	SE	*t*	*b*	SE	*t*	*b*	SE	*z*	*b*	SE	*z*
Critical Region															
Intercept	**1076**	**59**	**18.30**	**1583**	**72**	**22.10**	**1569**	**76**	**20.58**	**-0.94**	**0.10**	**-9.20**	**-089**	**0.09**	**-9.56**
Statement Phrasing	71	65	1.01	31	103	0.30	36	86	0.42	-0.12	0.12	-1.00	-0.05	0.13	-0.38
Degree of Imposition	11	37	0.30	-34	49	-0.68	-69	40	-1.73	-0.13	0.11	-1.15	-0.13	0.11	-1.15
Interaction	82	78	1.06	131	82	1.60	**169**	**80**	**2.10**	-0.11	0.22	-0.47	-0.08	0.22	-0.36

Stewart et al. (2018) Discourse Processes

Total Time



Stewart et al. (2018) Discourse Processes

- Suggests readers are rapidly sensitive to the degree of imposition associated with a request. Requests framed indirectly read more quickly with high degree of imposition than with low. Direct requests read at equivalent speed regardless of degree of imposition.

Indirect Replies

- People don’t like giving other people negative information that could be face threatening (Brown & Levinson, 1987).

- Face management argued to motivate the use of indirect replies.

- Indirect replies typically violate relevance (Holtgraves, 1998) - this violation triggers a search for a possible negative meaning.

- Face management is arguably quite a complex social process - are people sensitive to face management needs when reading conversations between two interlocutors?

Negative Situation.

Roberta and Andy are friends. Roberta is taking introductory chemistry this semester and is struggling on her course. Andy asked “How are you doing in chemistry?” She replied “The exams are not fair.” Andy planned to take the same course the following year. He was hopeful the course would be interesting.

Stewart et al. (2018) QJEP

Positive Situation.

Roberta and Andy are friends. Roberta is taking introductory chemistry this semester and is excelling on her course. Andy asked “How are you doing in chemistry?” She replied “The exams are not fair.” Andy planned to take the same course the following year. He was hopeful the course would be interesting.

Stewart et al. (2018) QJEP

Neutral Situation.

Roberta and Andy are friends. Roberta is taking introductory chemistry this semester that she attends on Tuesday afternoons. Andy asked “How are you doing in chemistry?” She replied “The exams are not fair.” Andy planned to take the same course the following year. He was hopeful the course would be interesting.

Stewart et al. (2018) QJEP

We manipulated whether the context was Negative, Positive, or Neutral.

This gives us a 1 factor with 3-levels repeated measure design.

- Twenty four participants.
- Twenty four experimental vignettes.
- Twenty four filler vignettes.
- Eye movements recorded using Eyelink 1000.

Stewart et al. (2018) QJEP

Two analysis regions:

She replied |**"The exams are not fair."**|$_{\text{critical}}$

|**Andy planned to take the same course the following year.** |$_{\text{post-critical}}$

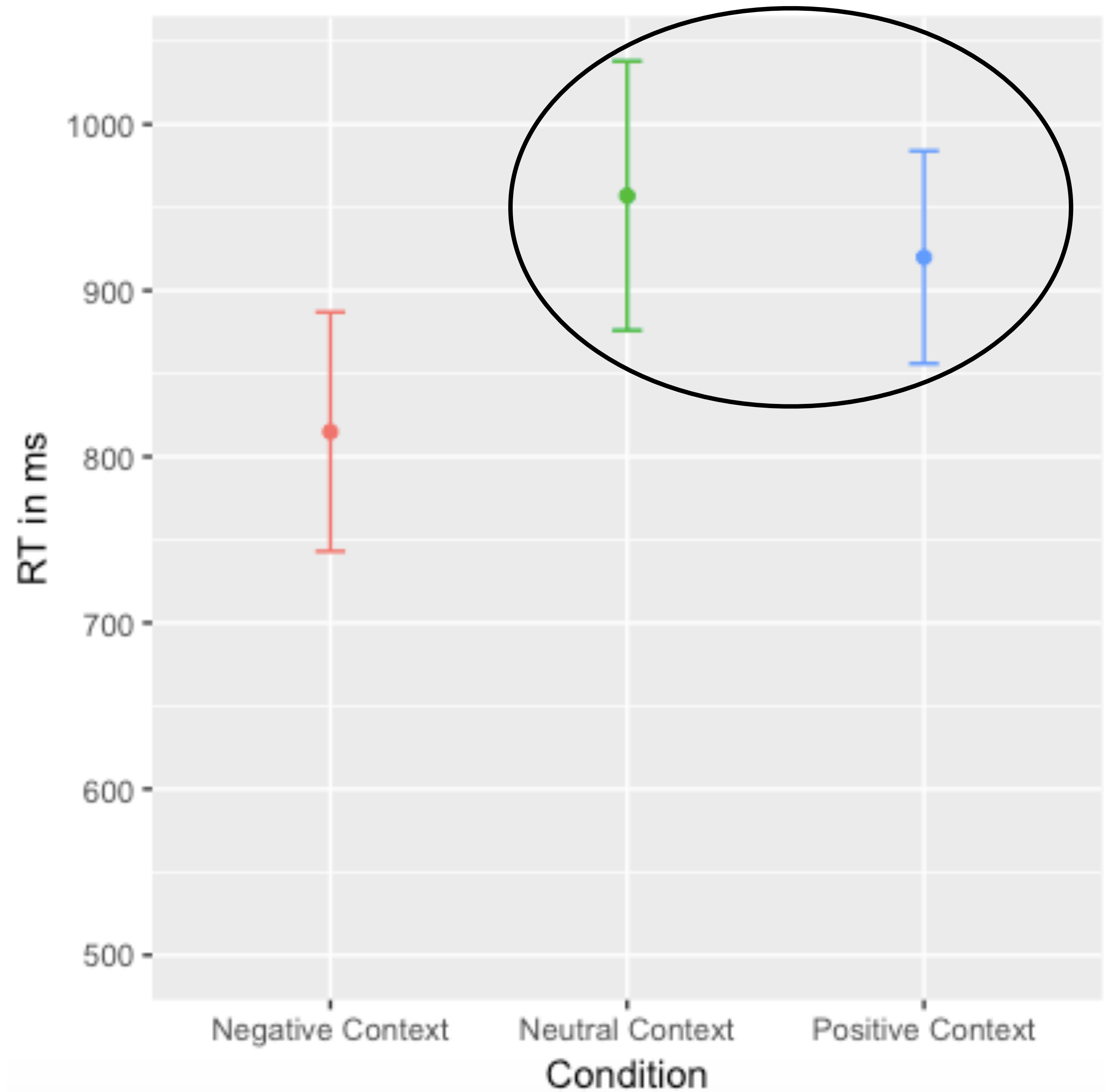
Stewart et al. (2018) QJEP

- Clear effects on Critical region:

	Duration measures								
	First Pass			Regression Path			Total Time		
	b	SE	*t*	*b*	SE	*t*	*b*	SE	*t*
Critical Region									
Intercept (Negative condition)	**815**	**72**	**11.39**	**1126**	**95**	**11.86**	**1086**	**83**	**13.07**
Positive condition	**104**	**45**	**2.33**	72	70	1.04	**121**	**51**	**2.38**
Neutral condition	**142**	**44**	**3.25**	99	72	1.36	**98**	**48**	**2.05**

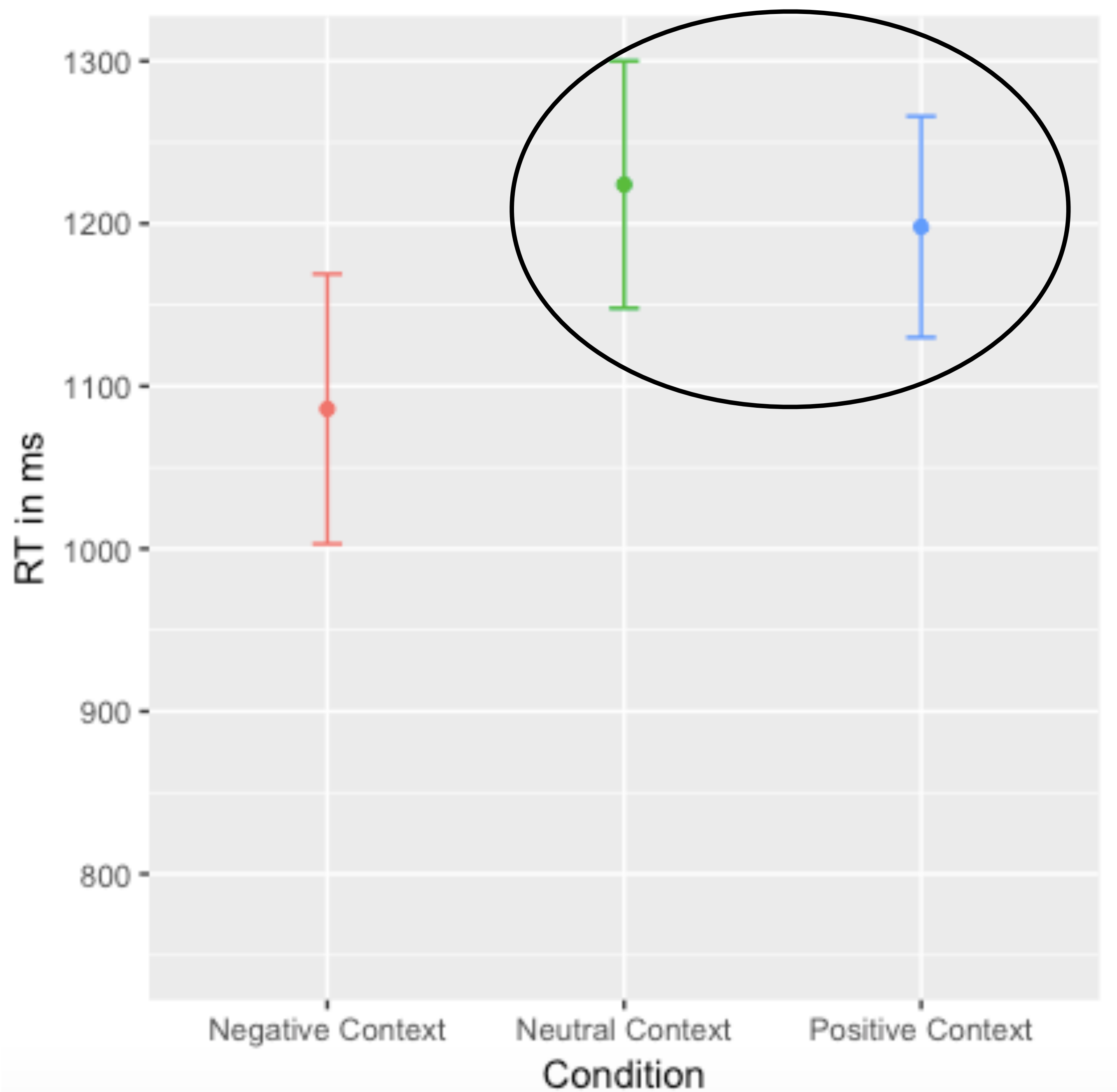
Stewart et al. (2018) QJEP

First Pass



Stewart et al. (2018) QJEP

Total Time



Stewart et al. (2018) QJEP

- Indirect replies (in the form of excuses) read most quickly in negative context (no both first pass and total time measures). These effects emerged on the critical region itself.

- On the post-critical region, disruption continued in the Positive, but not the Neutral context:

	Binomial measures					
	First Pass Regressions Out – by participants			First Pass Regressions Out – by items		
	b	SE	*z*	*b*	SE	*z*
Post-Critical Region						
Intercept (Negative condition)	**-3.565**	**0.628**	**-5.672**	**-2.856**	**0.361**	**-7.909**
Positive condition	**1.393**	**0.650**	**2.142**	**0.872**	**0.416**	**2.097**
Neutral condition	-0.143	0.860	-0.166	0.221	0.465	0.475

Stewart et al. (2018) QJEP

What does it all mean?

- Readers are **rapidly** sensitive to a range of factors during comprehension - including complex pragmatic information - and are able to quickly bring this knowledge into the frame to understand the implied and indirect meanings associated with both conditional and non-conditional statements, requests etc.

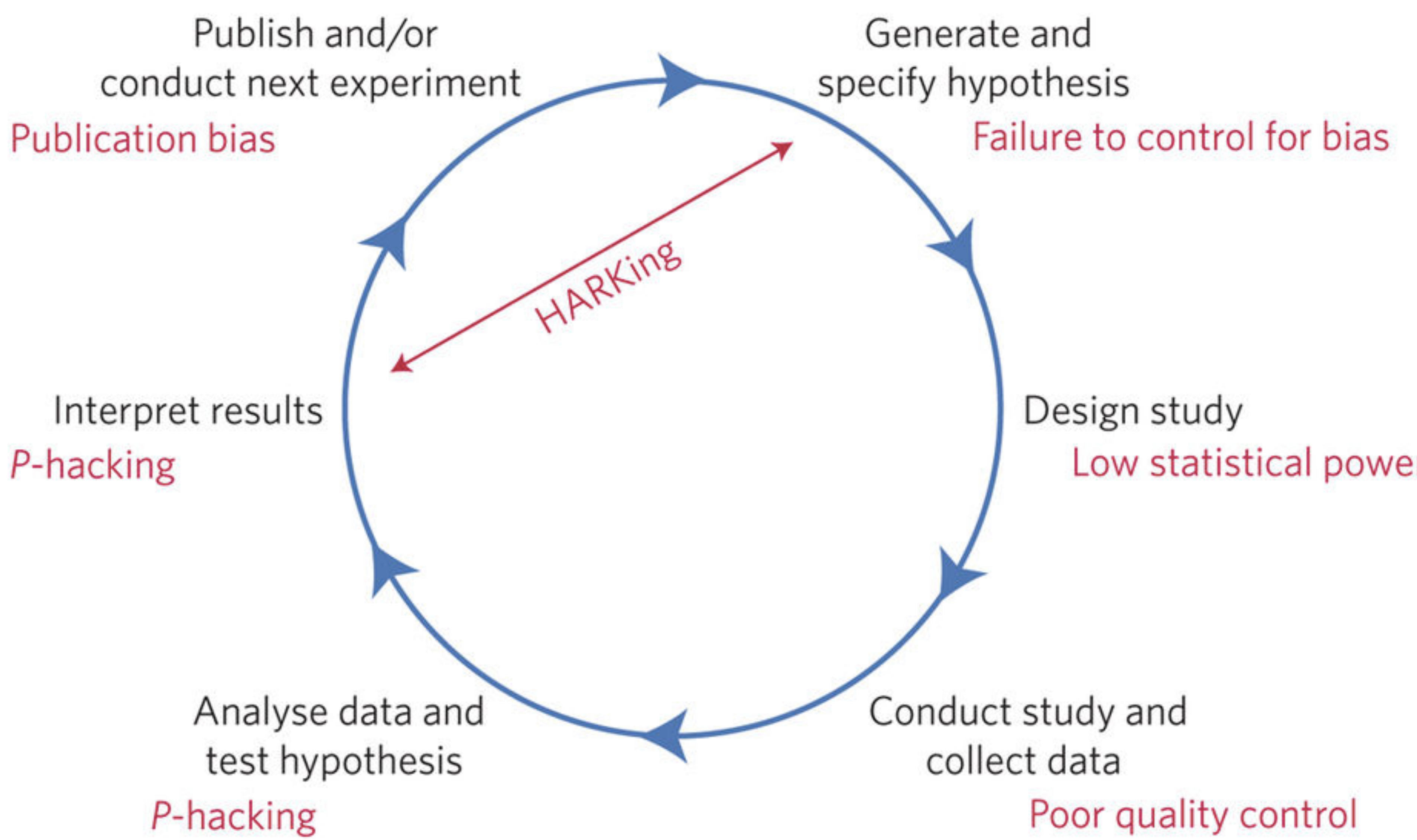
- Context is key.

But how would I carry out this research now?

- We now know that science has a replication problem.
- Ionnidis (2005), PLOS Medicine, most published research findings are false.
- Button et al. (2013), Nature Reviews Neuroscience, small sample size undermines the reliability of neuroscience.
- Baker (2015), Nature, 90% of scientists recognise a 'reproducibility crisis'.

Problems include p-hacking, lack of power, HARKing, failing (refusal) to share data and code, too many researcher degrees of freedom…

From: A manifesto for reproducible science

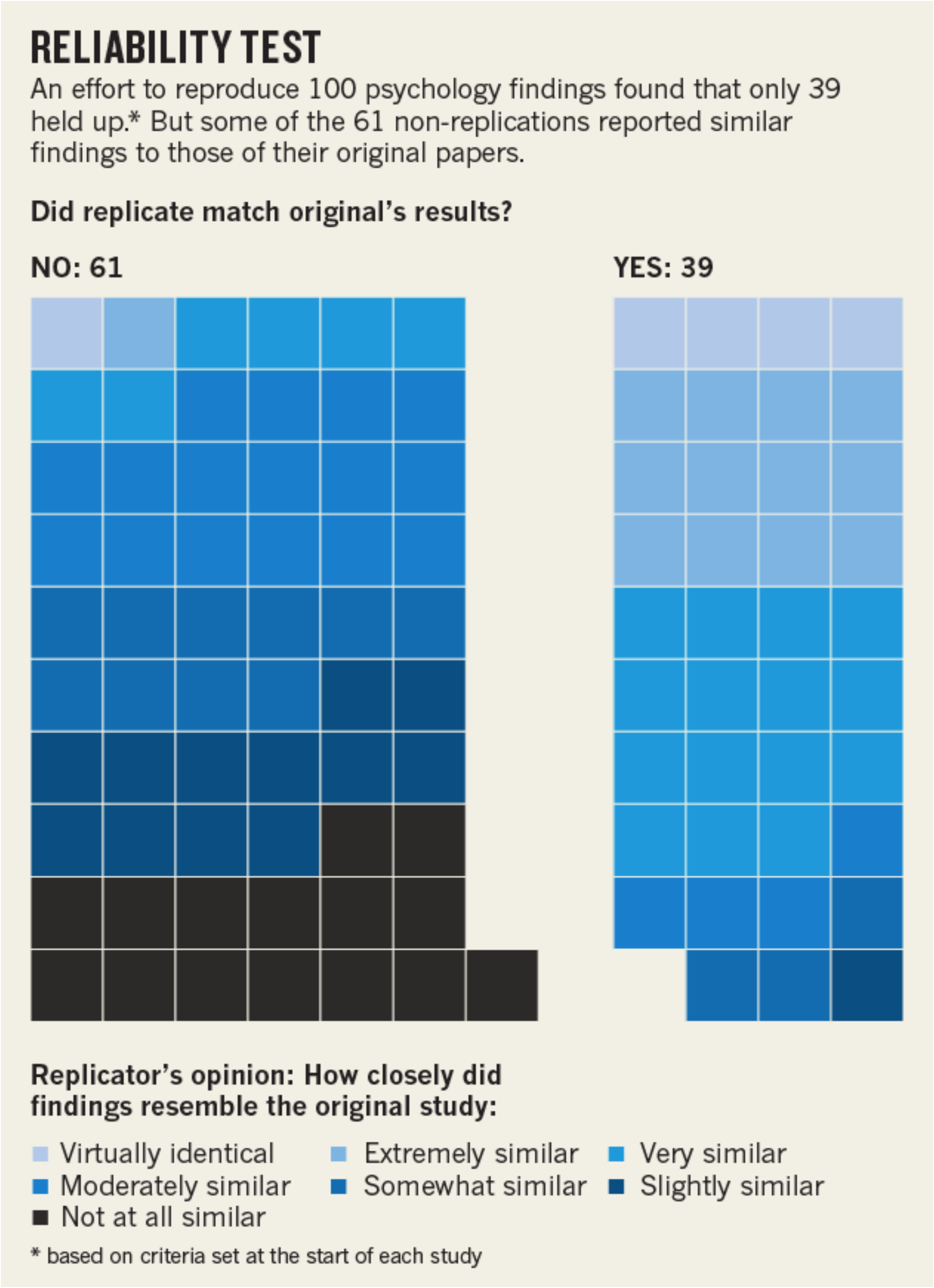


Munafo et al. (2017), *Nature Human Behaviour*

Why are so many studies not replicating?

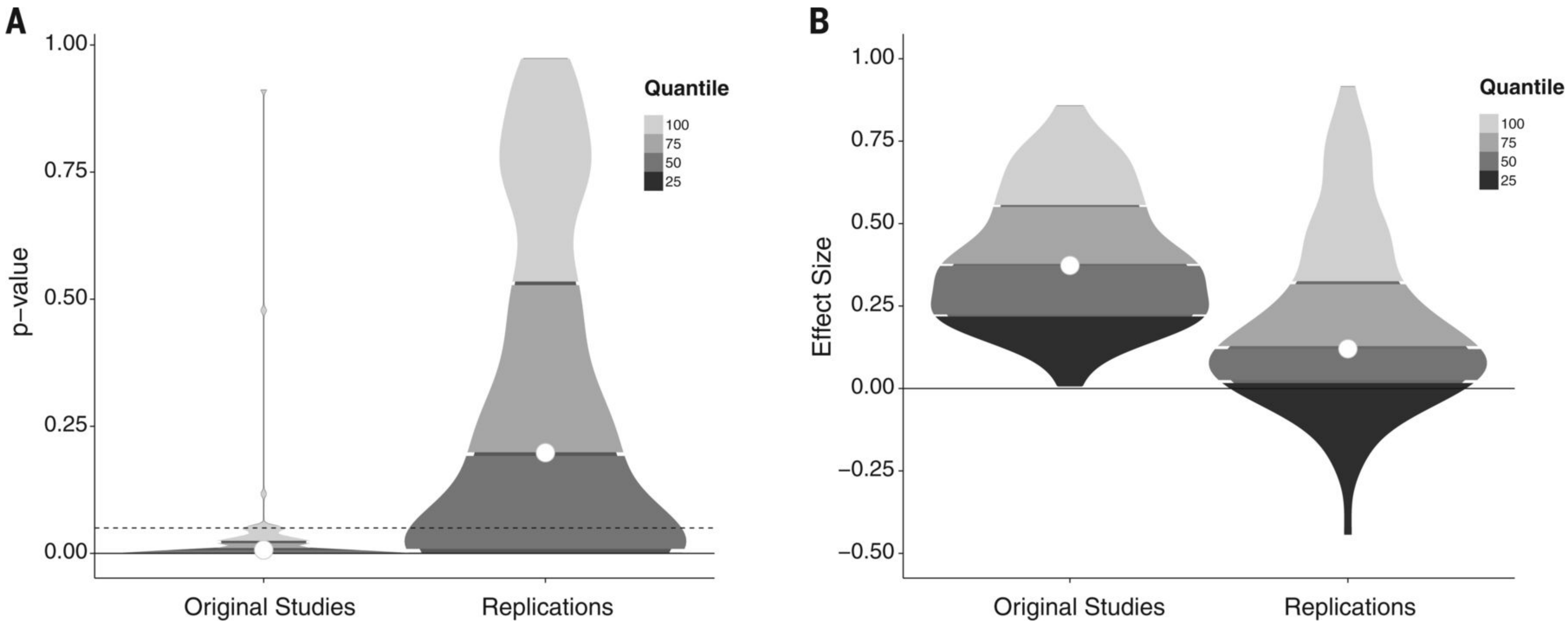
- There are too many studies with experimental power too low to detect the effect size of interest.
- One of the consequences of a low powered study is that when real effects are detected their magnitude is likely to be over-estimated.
- Studies which find the effect are published and studies that don't are not published - due to a bias to publish positive results.
- Future work may use the published effect size during *a priori* power analysis (and then fail to find the effect as the new study is effectively under-powered for what it's looking for).

Estimating the reproducibility of psychological science (Nosek et al., 2015)



270 authors tried to replicate 100 experiments drawn from high profile Psychology journals - *Psychological Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition.*

https://science.sciencemag.org/content/349/6251/aac4716



The *p*-values for the replication set formed a very different distribution to the *p*-values of the original studies. Similarly with the distribution of effect sizes.

https://science.sciencemag.org/content/349/6251/aac4716

ANNUAL REVIEWS

Annual Review of Psychology

Psychology's Renaissance

Leif D. Nelson,[1] Joseph Simmons,[2]
and Uri Simonsohn[2]

[1]Haas School of Business, University of California, Berkeley, California 94720;
email: Leif_Nelson@haas.berkeley.edu

[2]The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104;
email: jsimmo@upenn.edu, urisohn@gmail.com

"the overwhelming majority of published findings are statistically significant (Fanelli 2012, Greenwald 1975, Sterling 1959). On the other hand, the overwhelming majority of published studies are underpowered and, thus, theoretically unlikely to obtain results that are statistically significant."

Personality and Social Psychology Review
1998, Vol. 2, No. 3, 196–217

HARKing: Hypothesizing After the Results are Known

Norbert L. Kerr
Department of Psychology
Michigan State University

This article considers a practice in scientific communication termed HARKing (Hypothesizing After the Results are Known). HARKing is defined as presenting a post hoc hypothesis (i.e., one based on or informed by one's results) in one's research report as if it were, in fact, an a priori hypotheses. Several forms of HARKing are identified and survey data are presented that suggests that at least some forms of HARKing are widely practiced and widely seen as inappropriate. I identify several reasons why scientists might HARK. Then I discuss several reasons why scientists ought not to HARK. It is conceded that the question of whether HARKing's costs exceed its benefits is a complex one that ought to be addressed through research, open discussion, and debate. To help stimulate such discussion (and for those such as myself who suspect that HARKing's costs do exceed its benefits), I conclude the article with some suggestions for deterring HARKing.

WORLD VIEW *A personal take on events*

ROBERT TAYLOR

Rein in the four horsemen of irreproducibility

Dorothy Bishop *describes how threats to reproducibility, recognized but unaddressed for decades, might finally be brought under control.*

More than four decades into my scientific career, I find myself an outlier among academics of similar age and seniority: I strongly identify with the movement to make the practice of science more robust. It's not that my contemporaries are unconcerned about doing science well; it's just that many of them don't seem to recognize that there are serious problems with current practices. By contrast, I think that, in two decades, we will look back on the past 60 years — particularly in biomedical science — and marvel at how much time and money has been wasted on flawed research.

How can that be? We know how to formulate and test hypotheses in controlled experiments. We can account for unwanted variation with statistical techniques. We appreciate the need to replicate observations.

Yet many researchers persist in working in a way almost guaranteed not to deliver meaningful results. They ride with what I refer to as the four horsemen of the reproducibility apocalypse: publication bias, low statistical power, *P*-value hacking and HARKing (hypothesizing after results are known). My generation and the one before us have done little to rein these in.

In 1975, psychologist Anthony Greenwald noted that science is prejudiced against null hypotheses; we even refer to sound work supporting such conclusions as 'failed experiments.' This prejudice leads to publication bias: researchers are less likely to write up studies that show no effect, and journal editors are less likely to accept them. Consequently, no one can learn from them, and researchers waste time and resources

MANY RESEARCHERS PERSIST IN WORKING IN A WAY ALMOST **GUARANTEED** NOT TO DELIVER **MEANINGFUL RESULTS.**

be adequately powered. Other disciplines have yet to catch up.

I stumbled on the issue of *P*-hacking before the term existed. In the 1980s, I reviewed the literature on brain lateralization (how sides of the brain take on different functions) and developmental disorders, and I noticed that, although many studies described links between handedness and dyslexia, the definition of 'atypical handedness' changed from study to study — even within the same research group. I published a sarcastic note, including a simulation to show how easy it was to find an effect if you explored the data after collecting results (D. V. M. Bishop *J. Clin. Exp. Neuropsychol.* **12**, 812–816; 1990). I subsequently noticed similar phenomena in other fields: researchers try out many analyses but report only the ones that are 'statistically significant'.

This practice, now known as *P*-hacking, was once endemic to most branches of science that rely on *P* values to test significance of results, yet few people realized how seriously it could distort findings. That started to change in 2011, with an elegant, comic paper in which the authors crafted analyses to prove that listening to the Beatles could make undergraduates younger (J. P. Simmons *et al. Psychol. Sci.* **22**, 1359–1366; 2011). "Undisclosed flexibility," they wrote, "allows presenting anything as significant."

The term HARKing was coined in 1998 (N. L. Kerr *Pers. Soc. Psychol. Rev.* **2**, 196–217; 1998). Like *P*-hacking, it is so widespread that researchers assume it is good practice. They look at the data, pluck out a finding that looks exciting and write a paper to tell a story around this result. Of course, researchers should be free to explore their

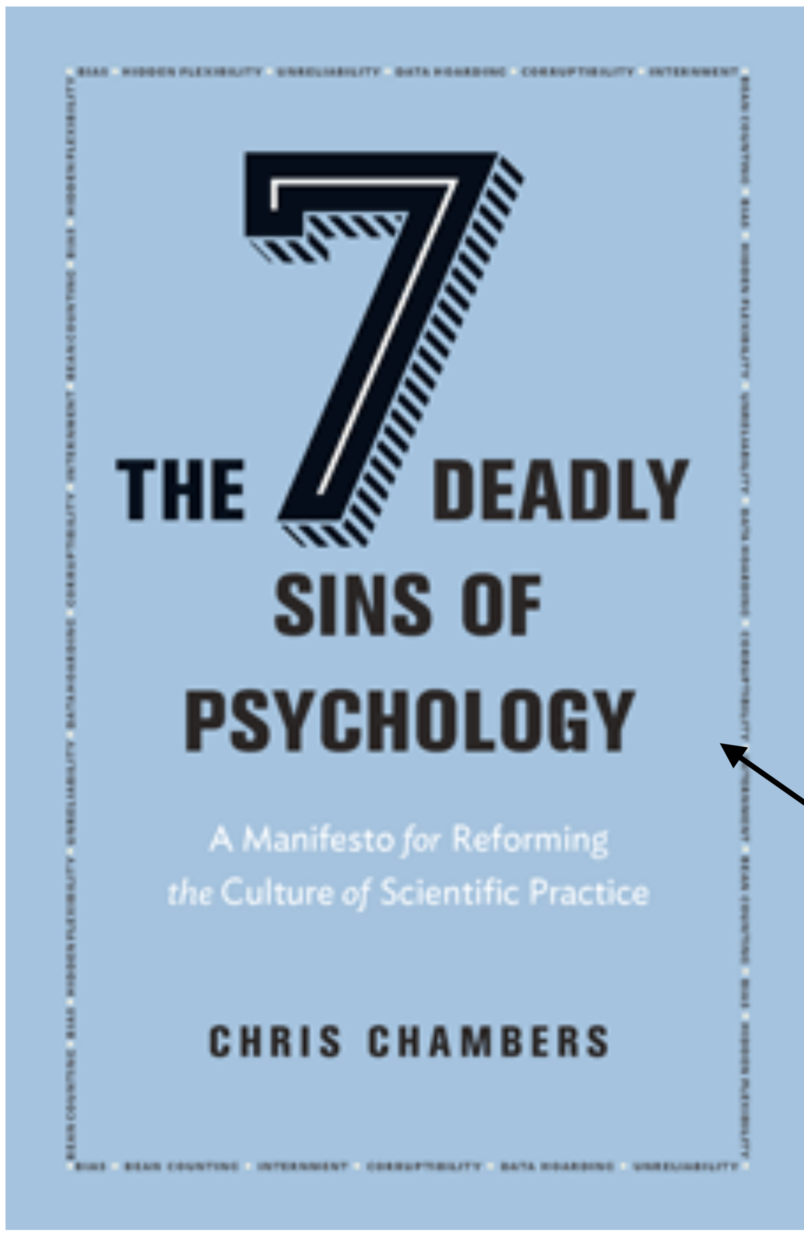
https://www.nature.com/articles/d41586-019-01307-2

Distinguishing between replicability and reproducibility (note, both are important!)

Replicable Research is when someone else can run a study the same as or conceptually equivalent to your one, and find a similar pattern of effects.

Reproducible Research is when someone else can take your data and your analysis code, run it and then find the same effects that you have reported.

A move towards open research…



Sins include *p*-hacking, lack of power, HARKing, failing (refusal) to share data and code, too many researcher degrees of freedom…

You really should read this book!

http://www.stat.columbia.edu/~gelman/

Andrew Gelman gives the following recommendations to researchers:

- Analyze all your data.
- Present all your comparisons.
- Make your data public.
- Put in the effort to take accurate measurements (low bias, low variance, and a large enough sample size).
- Do repeated-measures comparisons where possible.

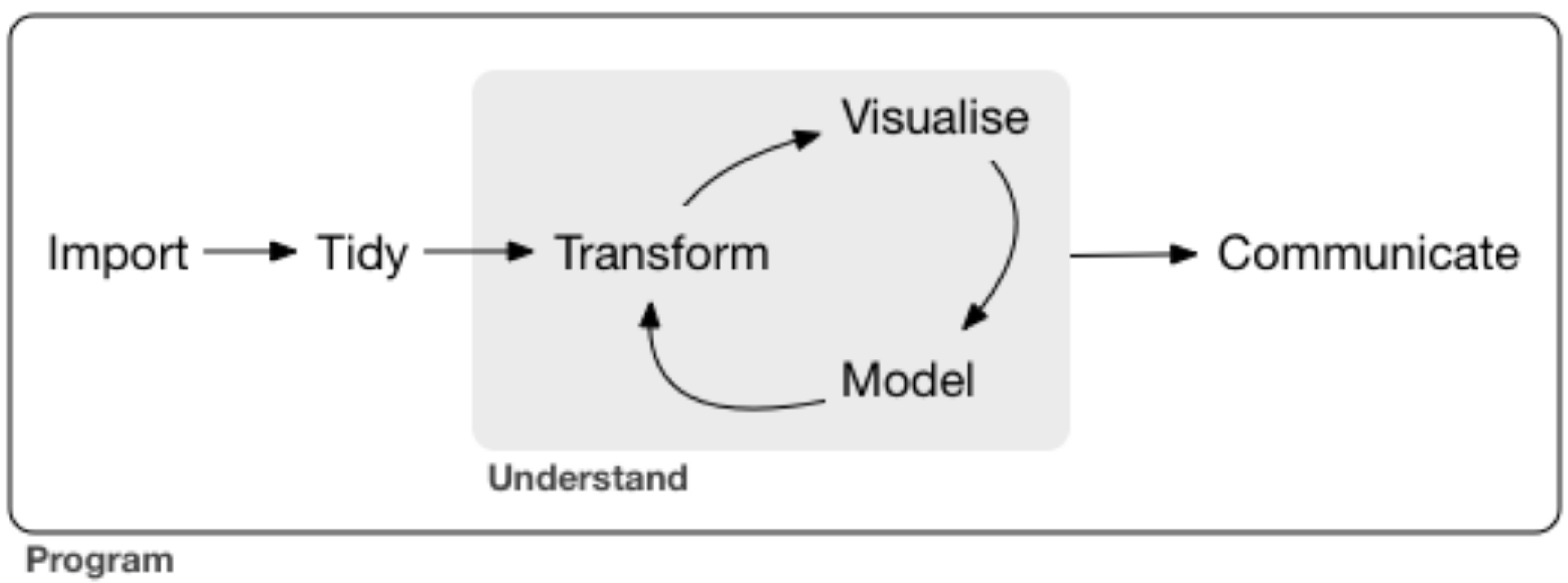
Before Data Collection

Open Research practices include…

- Specify your hypotheses and analysis plan.
- Pre-register your hypotheses and analysis plan at osf.io
- Conduct data simulation so that you can write your analysis script before you have your real data.
- Consider submitting as a registered report - more than 200 journals now support this route. This involves acceptance in principle before you have even started collecting your data.

After Data Collection

You need to use analysis software that allows for open sharing and reproducibility of the entire data wrangling/analysis/write-up workflow.



Hadley Wickham and Garrett Grolemund

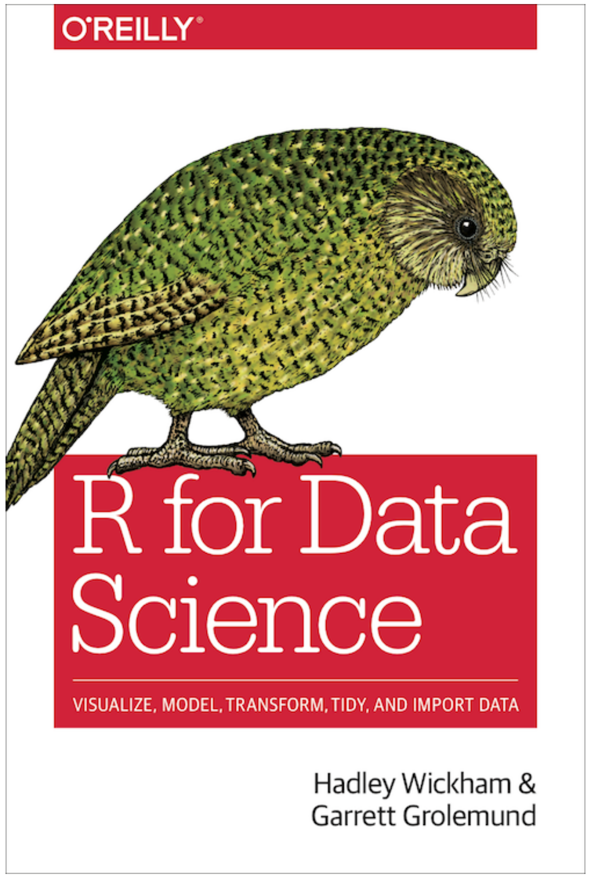
Use R for Data Analysis

Image credit Darren Dahly @statsepi

If statistics programs/languages were cars...

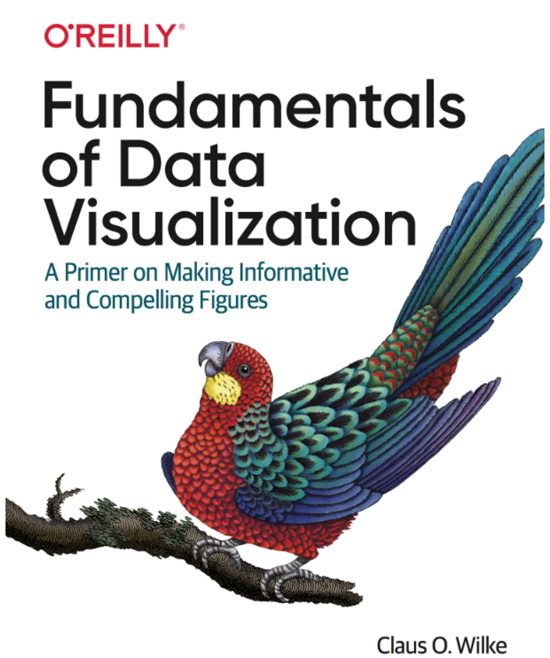
What role can R play in Open Research?

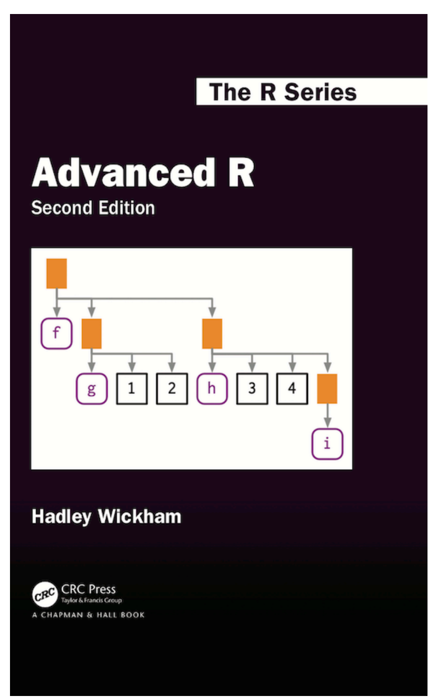
- R scripts are easy to share allowing for reproducibility and easy public sharing of data and code.
- R is free, open source software that is much more flexible and powerful than SPSS.
- There is an active R community continuously updating statistical tests and packages that run in R.
- As R is a programming language, it forces you to <u>know</u> your data.

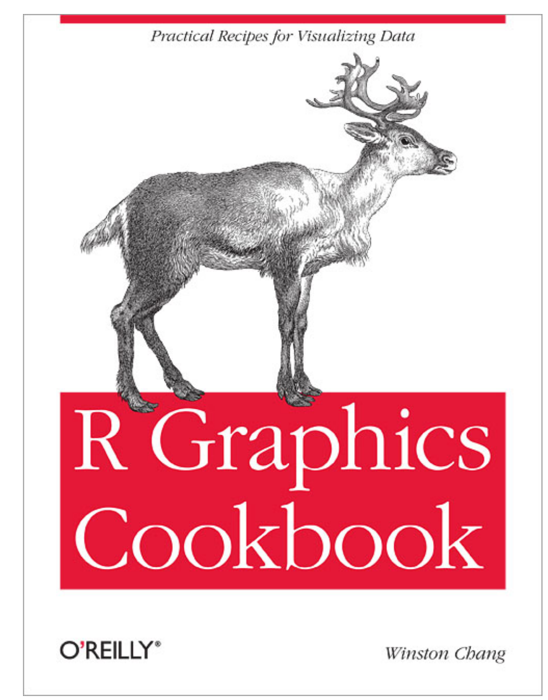


Available electronically for free at:

http://r4ds.had.co.nz







You can share your data at osf.io or on GitHub:

ajstewartlang / **Comprehension-of-indirect-requests-is-influenced-by-their-degree-of-imposition**

Watch 0 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Branch: master | **Comprehension-of-indirect-requests-is-influenced-by-their-degree-of-imposition** / **RPs.csv** | Find file | Copy path

ajstewartlang Made consistent the labelling of factors in data files and in paper | 7b3b3b1 on 29 Mar 2017

0 contributors

1681 lines (1681 sloc) | 69.7 KB | Raw | Blame | History

Search this file...

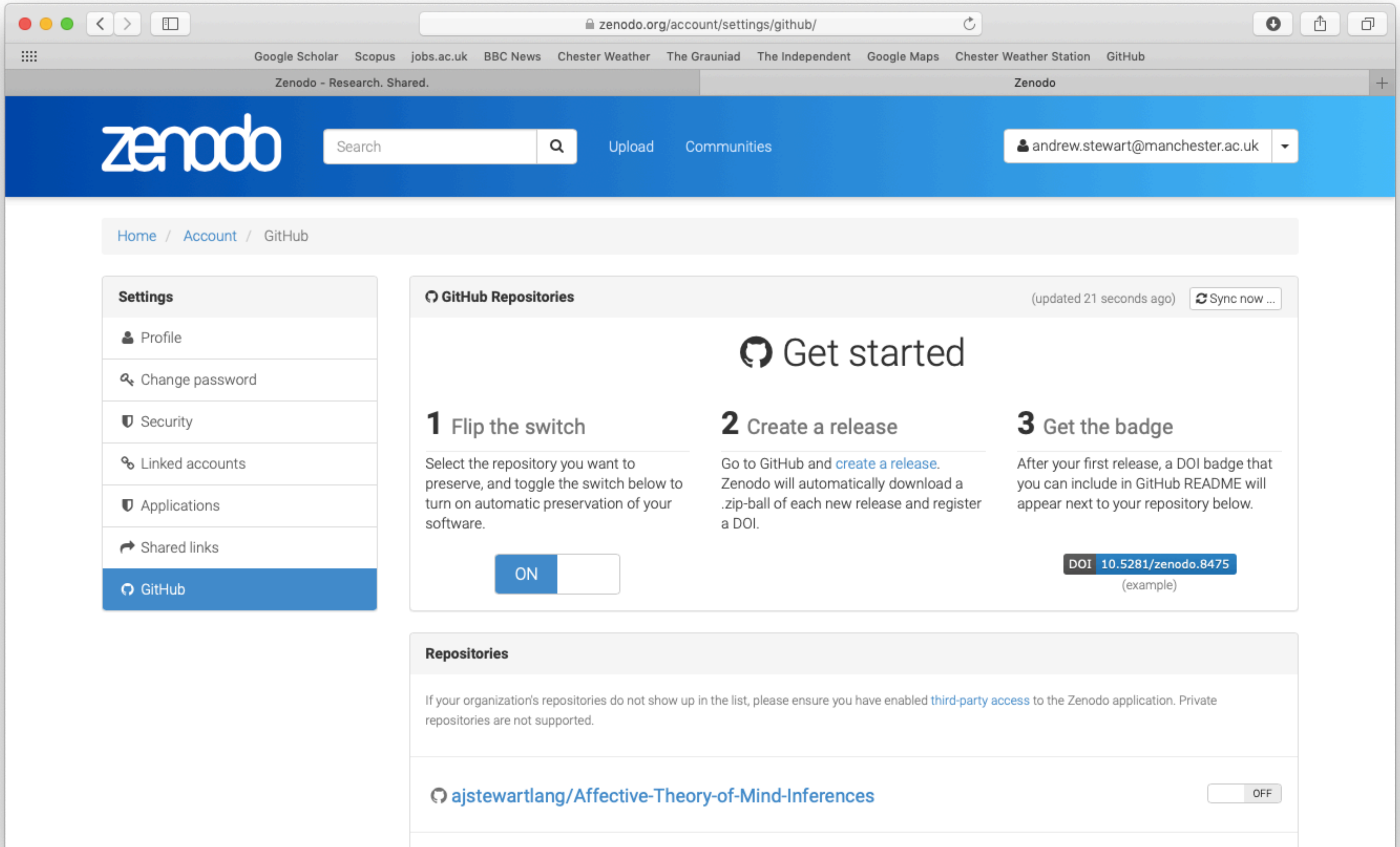
P.s	Item	Condition	Probmanip	Speaker	statement	response	final	Meaning	Imposition
1	1	1	1708	302	1399	1867	1206	Indirect	High
1	2	2	1466	296	1377	1674	828	Indirect	Low
1	3	3	1393		1494	1950	1812	Direct	High
1	4	4	2463	530	1691	1866	965	Direct	Low
1	5	1	1552	267	1332	1477	1345	Indirect	High
1	6	2	1445	444	1004	1067	797	Indirect	Low
1	7	3	2159	501	739	1231	2240	Direct	High
1	8	4	1459		1086	946	978	Direct	Low
1	9	1	3302		1503	900	1736	Indirect	High

FaceTime

alongside your analysis code:

```r
FPs$Meaning <- as.factor(FPs$Meaning)
FPs$Imposition <- as.factor(FPs$Imposition)

#this sets up the contrasts so that the intercept in the mixed LMM is the grand mean (i.e., the mean of all conditions)
my.coding <- matrix (c(.5, -.5))

contrasts (FPs$Meaning) <- my.coding
contrasts (FPs$Imposition) <- my.coding

#construct the models with crossed random effects for subjects and items for the pre-critical, critical and post-crtical region
fpmodelprec <- lmer (Probmanip ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning+Imposition |Item), data=FPs, REM
summary (fpmodelprec)
lsmeans (fpmodelprec, pairwise~Meaning*Imposition, adjust="none")

fpmodelc <- lmer (statement ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning*Imposition |Item), data=FPs, REML=T
summary (fpmodelc)
lsmeans (fpmodelc, pairwise~Meaning*Imposition, adjust="none")

fpmodelpostc <- lmer (response ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning+Imposition |Item), data=FPs, REM
summary (fpmodelpostc)
lsmeans (fpmodelpostc, pairwise~Meaning*Imposition, adjust="none")

#Regression Path Analysis
#Read in Regression Path data
RPs <- read.csv("~/RPs.csv")

RPs$Meaning <- as.factor(RPs$Meaning)
RPs$Imposition <- as.factor(RPs$Imposition)

contrasts (RPs$Meaning) <- my.coding
contrasts (RPs$Imposition) <- my.coding

#construct the models with crossed random effects for subjects and items for the pre-critical, critical and post-crtical region
rpmodelprec <- lmer (Probmanip ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning*Imposition |Item), data=RPs, REM
```

And make it citable with a DOI via Zenodo:



Sharing your computational environment

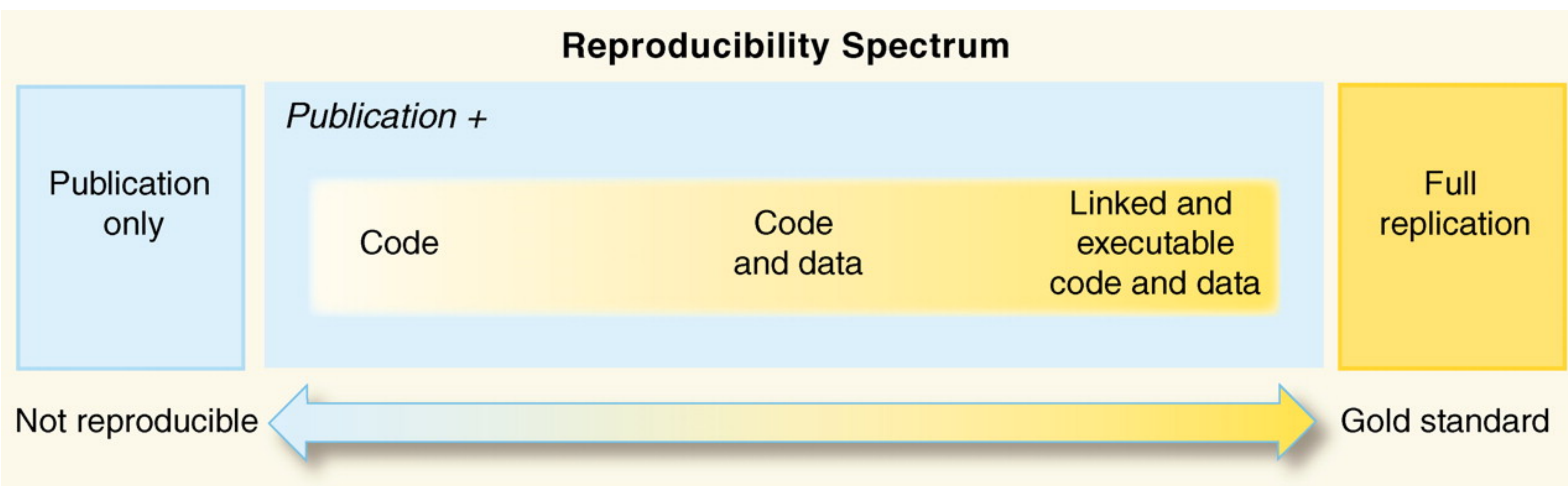
PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

+ See all authors and affiliations

Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847



Join our Open Research Working Group

- Open Research Working Group at Manchester founded in November 2018 by myself and Caroline Jay (Computer Science) - subscribe to our listserve:

https://listserv.manchester.ac.uk/cgi-bin/wa?REPORT=OPEN_RESEARCH

- Lots of OS activities incl. reproducibility journal club (ReproducibiliTea) meetings.

- Check out the Network of Open Research Working groups: https://osf.io/vgt3x/

North West Open Research Hub

- We are part of a broader network in the NW including Lancaster, Keele, MMU, Leeds, Chester, Sheffield.

- We are also part of the UK Reproducibility Network funded/supported by UKRI, research England, MRC, NERC, ESRC, Wellcome, Universities UK, JISC, British Neuroscience Association (amongst others).

- Links to Project Tier, The Carpentries, Software Sustainability Institute, The Turing Way etc.

The UKRN

University of BRISTOL

UK Reproducibility Network

Current students Current staff Alumni

search

UK Reproducibility Network

- About
- Local Networks
- Steering Group
- Advisory Board
- Stakeholders
- Contacts
- PhD Students
- UKRN Events

The UK Reproducibility Network (UKRN)

The UK Reproducibility Network (UKRN) is a peer-led consortium that aims to ensure the UK retains its place as a centre for world-leading research.

This will be done by investigating the factors that contribute to robust research, promoting training activities, and disseminating best practice, and working with stakeholders to ensure coordination of efforts across the sector.

It is led by Marcus Munafò (Bristol), Chris Chambers (Cardiff), Laura Fortunato (Oxford), Alexandra Collins (Imperial), and Malcolm Macleod (Edinburgh).

UKRN works across disciplines, ranging from the arts and humanities to the physical sciences, with a particular focus on the biomedical sciences.

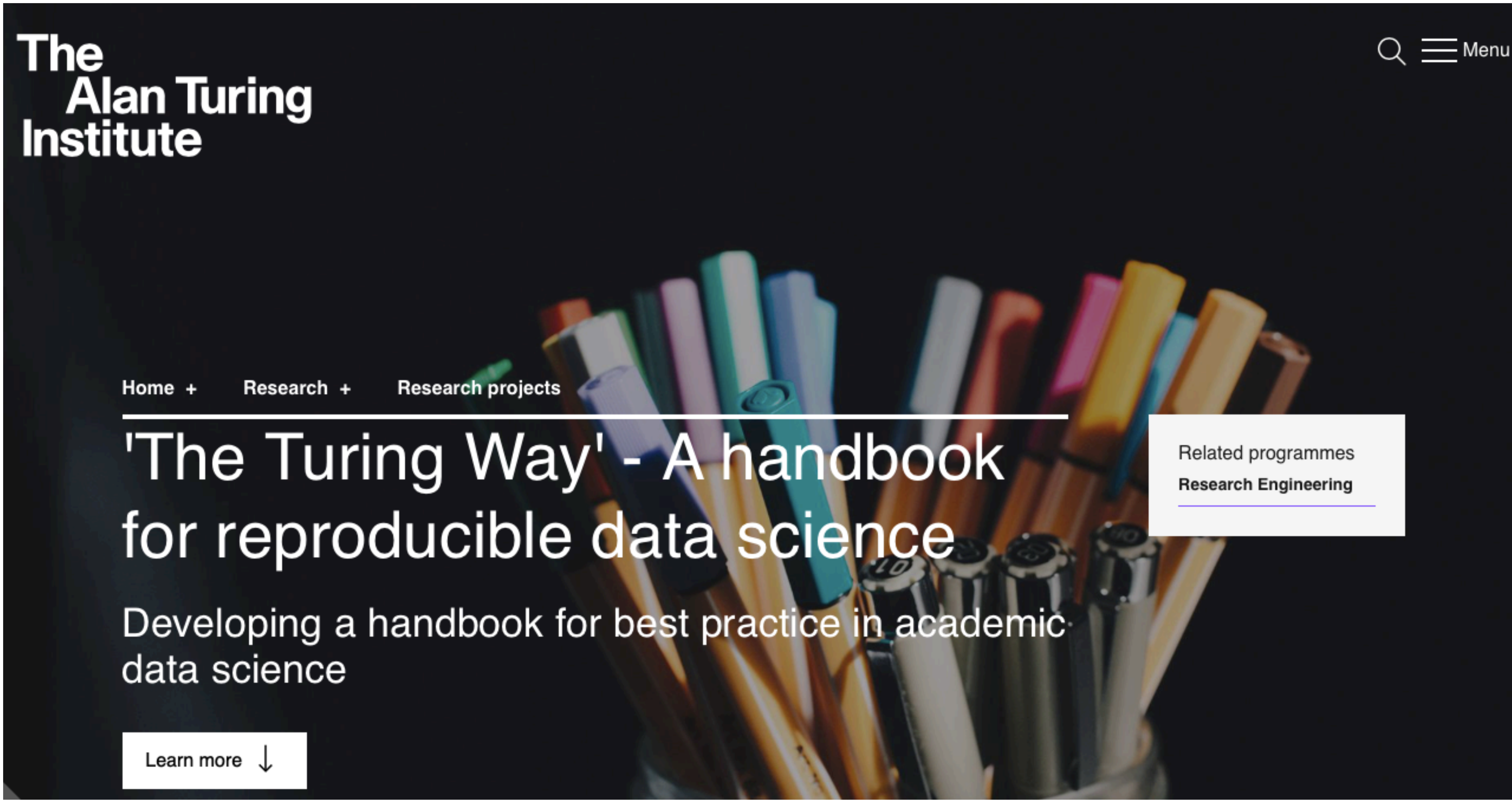
UKRN News

See the latest news about the Network

UK RN

http://www.bristol.ac.uk/psychology/research/ukrn/

The Turing Way



https://www.turing.ac.uk/research/research-projects/turing-way-handbook-reproducible-data-science

The Software Sustainability Institute

Software
Sustainability
Institute

About   Programmes and Events

Reproducible research

The reproducibility of research is at the very heart of the scientific method. As more research is based on results that are generated by software, there must be an increased focus on developing software that is reliable and which can be easily proven to produce reproducible results.

https://www.software.ac.uk/about/manifesto