What did James Heathers argue about the rate of fake studies?

StatCheck

Ian Hussey

Digitalisation of Psychology

What does StatCheck do?

What did you find in your articles?

Results across all your articles

error	n_tests	percent
no	14	3.2
other	427	96.8

any_error	n_articles	percent
FALSE	121	89.6
TRUE	14	10.4

But zero decision errors!

What could have gone wrong with this analysis?

What assumptions or known flaws are there?

Understanding *p* values

Ian Hussey

Digitalisation of Psychology

What is a *p* value?

How common are significant p values when the null hypothesis is true?

How common are significant p values when the alternative hypothesis is true?

	$p < .05$	$p > .05$
True real effect	True positive result	False negative result
True null effect	False positive result	True negative result

I find $p = .078$
false negative or true negative?

	$p < .05$	$p > .05$
True real effect	True positive result	False negative result
True null effect	False positive result	True negative result

I find $p = .078$

false negative or true negative?

	$p < .05$	$p > .05$
True real effect	True positive result	False negative result
True null effect	False positive result	True negative result

UNKNOWABLE

I find $p = .078$
false negative or true negative?

	$p < .05$	$p > .05$
True real effect	True positive result	False negative result
True null effect	False positive result ***alpha value = 0.05 = 5%***	True negative result

I find $p = .022$
false positive or true positive?

	$p < .05$	$p > .05$
True real effect	True positive result	False negative result
True null effect	False positive result	True negative result

I find $p = .022$

false positive or true positive?

	$p < .05$	$p > .05$
True real effect	True positive result	False negative result
True null effect	False positive result	True negative result

UNKNOWABLE

I find $p = .022$
false positive or true positive?

	$p < .05$	$p > .05$
True real effect	True positive result ***Statistical power %***	False negative result ***Beta % (1 - power)***
True null effect	False positive result	True negative result

What most people want to know:

Given my data,
how probable is it my hypothesis is true?

Given the null hypothesis is true, how improbable is my data?

If very improbable, lets act as if my hypothesis is true

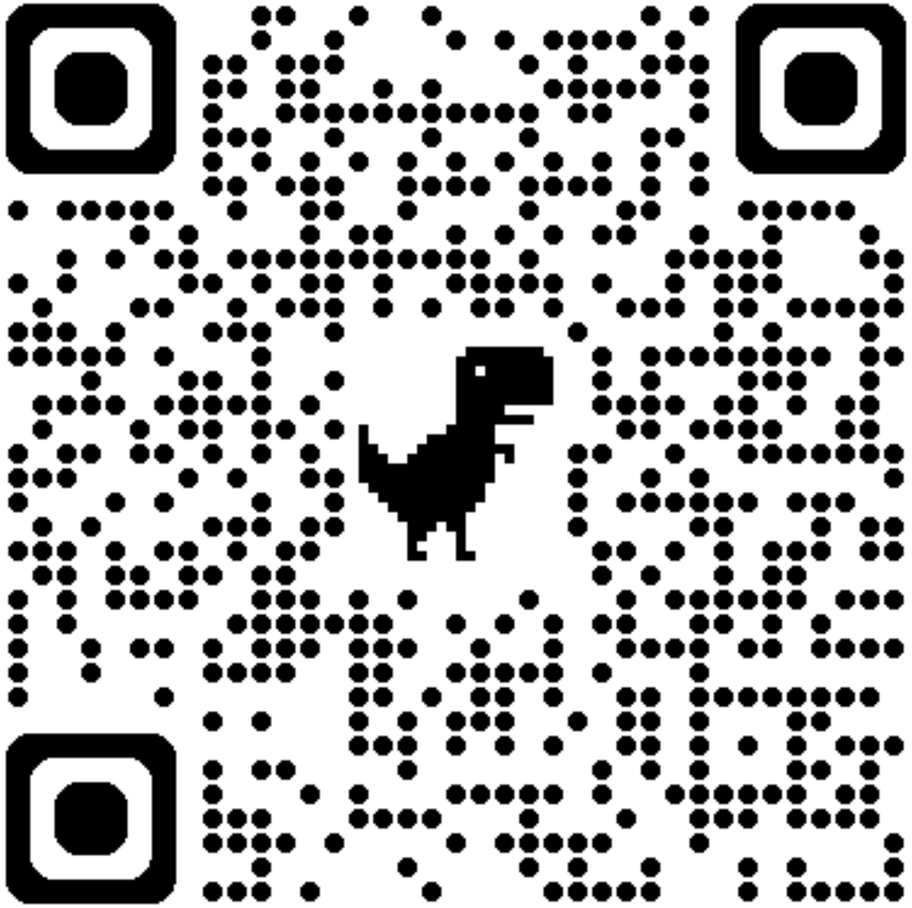
1. **This is extremely non-intuitive**
2. **This is a p-value**

We can say very little about the trustworthiness of individual *p* values without lots of information about the study.

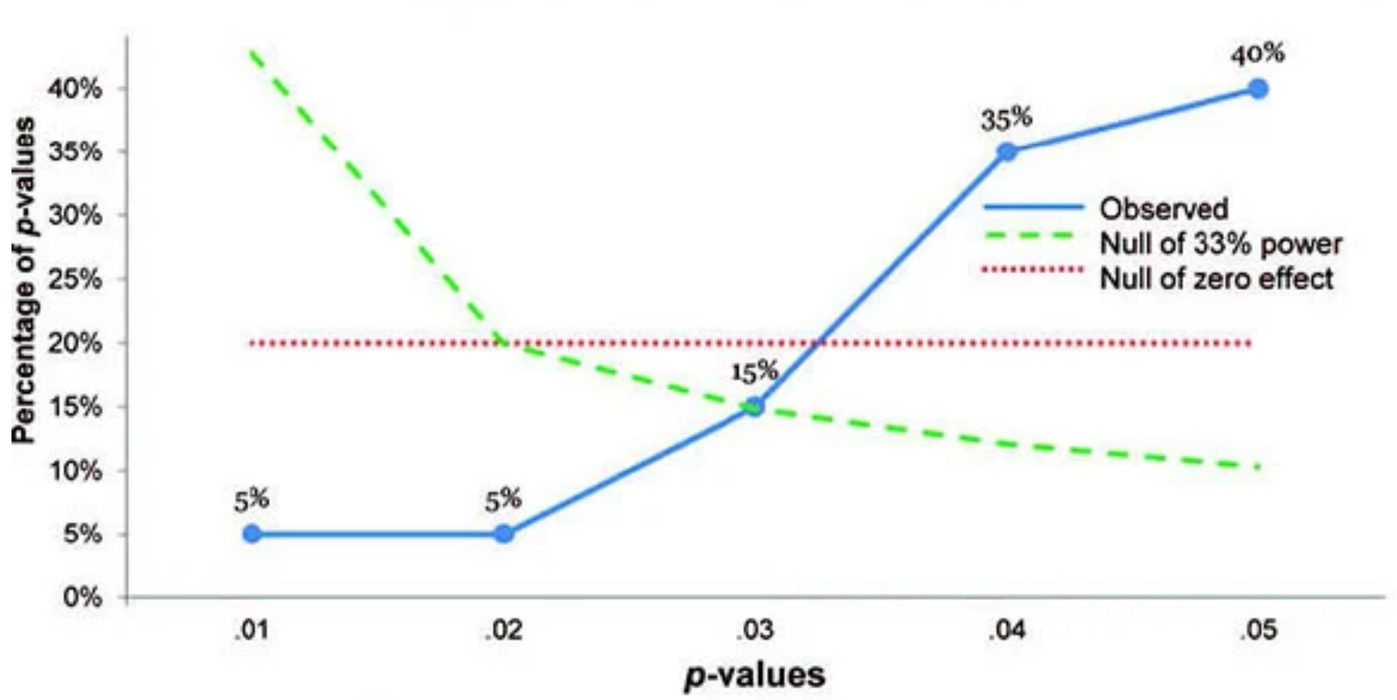
But there are useful patterns among large numbers of *p* values

What is the distribution of (unbiased) *p* values?

Understanding the distribution of unbiased *p* values



What could cause this distribution of p-values (in blue)?

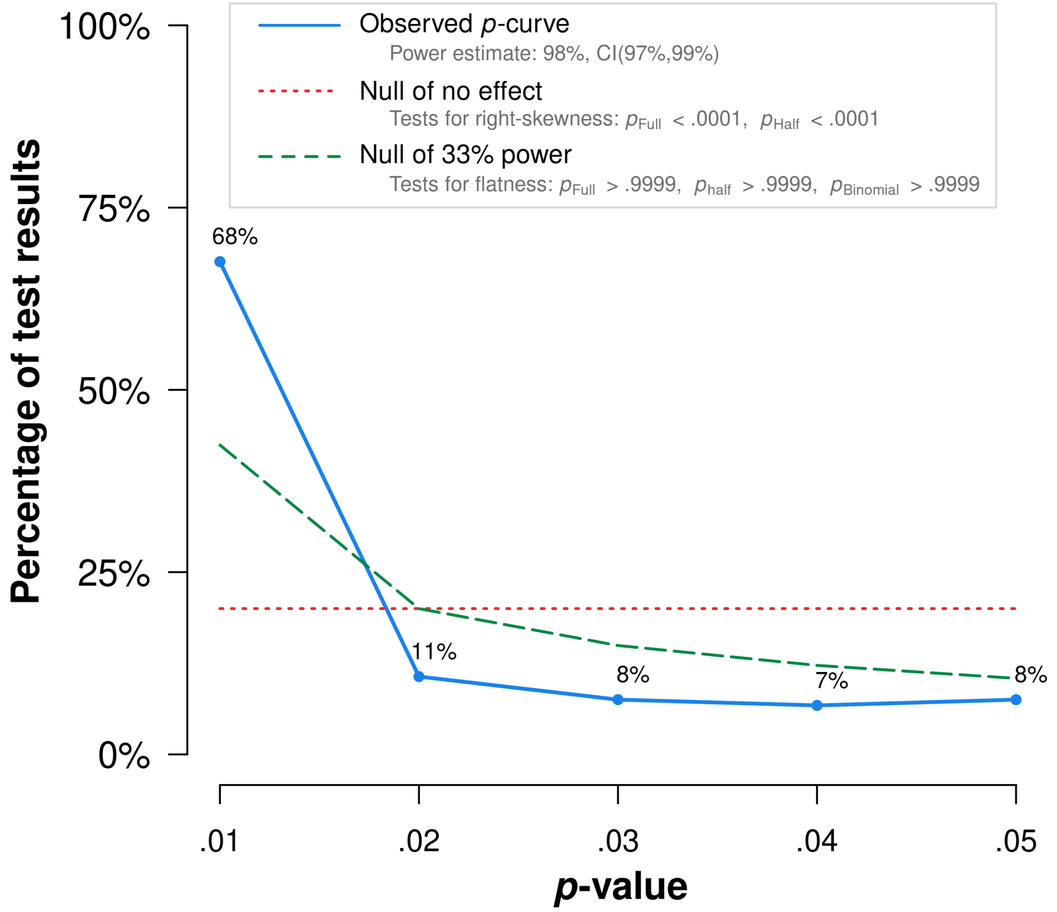


p-curve

https://www.p-curve.com/

Simonsohn, Simmons, & Nelson (2016)

Distribution of p values you extracted in your assignment



p-curve
https://www.p-curve.com/
Simonsohn, Simmons, & Nelson (2016)

(Mis)understand non-significant *p* values

October 1981
Final Report

HS-806-182

U.S. Department of Transportation
National Highway Traffic Safety Administration

The Effect of Right-Turn-On-Red on Pedestrian and Bicyclist Accidents

D. F. Preusser
W. A. Leaf
K. B. DeBartolo
R. D. Blomberg

Dunlap and Associates, Inc.
One Parkland Drive
Darien, Connecticut 06820

Contract No. DOT-HS-6-01411
Contract Amount $146,727

Pedestrians in New York Upstate Urban, Signalized Location with Vehicle Turning Right.

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Avg.
1974	6	4	5	3	3	4	0	3	2	7	6	3	3.833
1975	8	3	9	4	3	3	2	1	4	5	5	6	4.417
1976	2	3	2	2	6	1	2	2	3	6	4	7	3.333
1977	4	2	9	10	5	3	0	3	3	3	12	4	4.833
1978	4	5	5	7	8	3	3	4	4	11	9	10	6.083
Avg.	4.8	3.4	6.0	5.2	5.0	2.8	1.4	2.6	3.2	6.4	7.2	6.0	4.500

Analysis of Variance

Source	Mean Square	d.f.	F
Year	13.292	4	2.959
Month	16.018	11	3.566
Yr x Mon	4.492	44	

Time Series Analysis

Model	SE residual	Q	d.f.	p
None	2.690	44.53	25	.011
$(1-B^{12})$	3.306	68.36	25	.000
Pre-RTOR	1.890	20.31	22	n.s.
Intervention	2.398	23.30	22	[illegible]
Hypothesis	2.267	27.55	22	.20

Descriptive Models

Pre-RTOR: (36 months) $(1 - .485B^{12})(Y_t - 3.758) = (1 - .383B^{4})a_t$

Intervention: $(1 - .240^{*}B^{12})(Y_t - 3.854) = 1.467X_t + (1 - .433B^{4})a_t$

Hypothesis Model (Intervention)

$$Y_t = 1.567X_t + \frac{(1 - .212^{*}B^{4})(1 - .898B^{12})}{(1 - B^{12})}a_t$$

*parameter not significant

Right turn on red

29 extra deaths (+9.4%)

Pedestrians in New York Upstate Urban,
Signalized Location with Vehicle Turning Right.

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Avg.
1974	6	4	5	3	3	4	0	3	2	7	6	3	3.833
1975	8	3	9	4	3	3	2	1	4	5	5	6	4.417
1976	2	3	2	2	6	1	2	2	3	6	4	7	3.333
1977	4	2	9	10	5	3	0	3	3	3	12	4	4.833
1978	4	5	5	7	8	3	3	4	4	11	9	10	6.083
Avg.	4.8	3.4	6.0	5.2	5.0	2.8	1.4	2.6	3.2	6.4	7.2	6.0	4.500

Analysis of Variance

Source	Mean Square	d.f.	F
Year	13.292	4	2.959
Month	16.018	11	3.566
Yr x Mon	4.492	44	

Time Series Analysis

Model	SE residual	Q	d.f.	p
None	2.690	44.53	25	.011
$(1-B^{12})$	3.306	68.36	25	.000
Pre-RTOR	1.890	20.31	22	n.s.
Intervention	2.398	23.30	22	[illegible]
Hypothesis	2.267	27.55	22	.20

Descriptive Models

Pre-RTOR: (36 months) $(1 - .485B^{12})(Y_t - 3.758) = (1 - .383B^{4})a_t$

Intervention: $(1 - .240^{*}B^{12})(Y_t - 3.854) = 1.467X_t + (1 - .433B^{4})a_t$

Hypothesis Model (Intervention)

$$Y_t = 1.567X_t + \frac{(1 - .212^{*}B^{4})(1 - .898B^{12})}{(1 - B^{12})}a_t$$

*parameter not significant

Intention Invention and the Affect Misattribution Procedure: Reply to Bar-Anan and Nosek (2012)

Personality and Social
Psychology Bulletin
39(3) 375–386

DOI: 10.1177/0146167212475225
pspb.sagepub.com
SAGE

B. Keith Payne[1], Jazmin Brown-Iannuzzi[1], Melissa Burkley[2], Nathan L. Arbuckle[3], Erin Cooley[1], C. Daryl Cameron[1], and Kristjen B. Lundberg[1]

Abstract

A recent study of the affect misattribution procedure (AMP) found that participants who retrospectively reported that they intentionally rated the primes showed larger effect sizes and higher reliability. The study concluded that the AMP's validity depends on intentionally rating the primes. We evaluated this conclusion in three experiments. First, larger effect sizes and higher reliability were associated with (incoherent) retrospective reports of both (a) intentionally rating the primes and (b) being unintentionally influenced by the primes. A second experiment manipulated intentions to rate the primes versus targets and found that this manipulation produced systematically different effects. Experiment 3 found that giving participants an option to "pass" when they felt they were influenced by primes did not reduce priming. Experimental manipulations, rather than retrospective self-reports, suggested that participants make post hoc confabulations to explain their responses. There was no evidence that validity in the AMP depends on intentionally rating primes.

Keywords

affect misattribution procedure, implicit social cognition, implicit attitudes

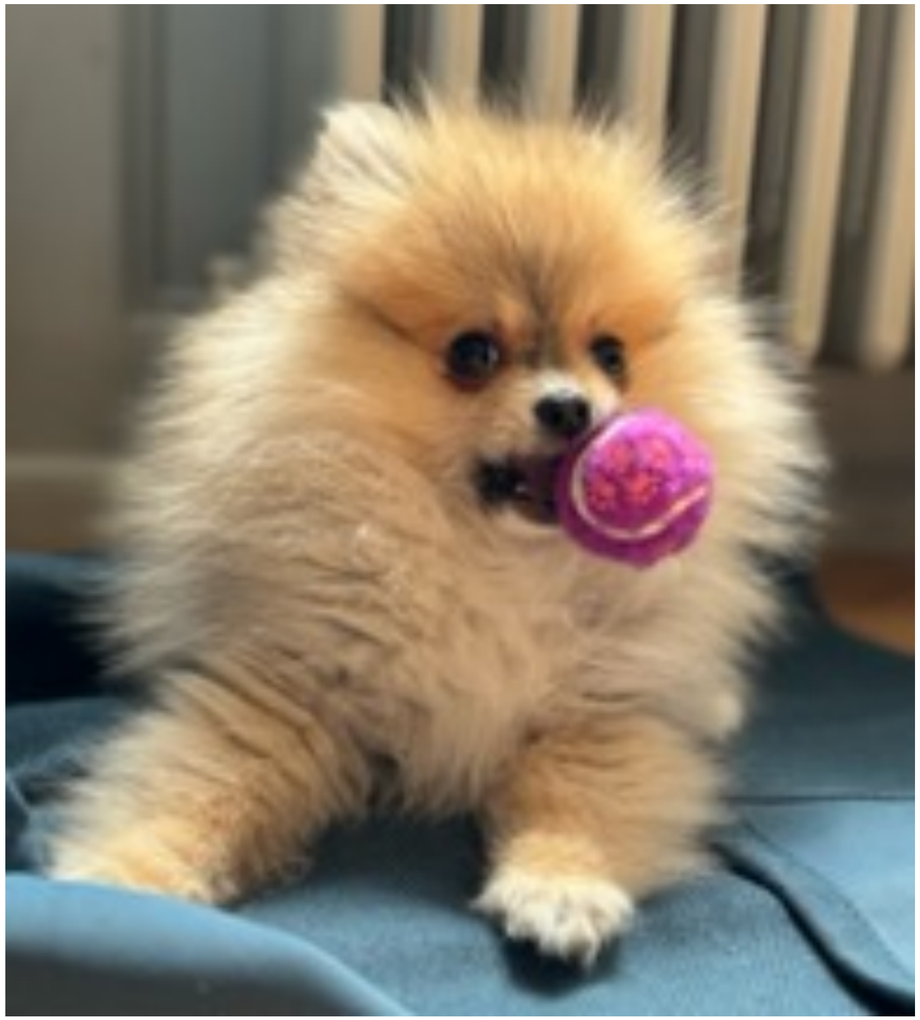
Received August 23, 2012; revision accepted November 4, 2012

Misinterpretations of non-significant *p* values are very common (Aczel et al., 2018)

“Absence of evidence does not equal evidence of absence”

“Abwesenheit von Evidenz für einen Effekt ist nicht gleichbedeutend mit Evidenz für die Abwesenheit eines Effekts”

Non-significant *p* values are not evidence of zero effect



Rummy and the 💩 alarm

Rummy is our puppy.

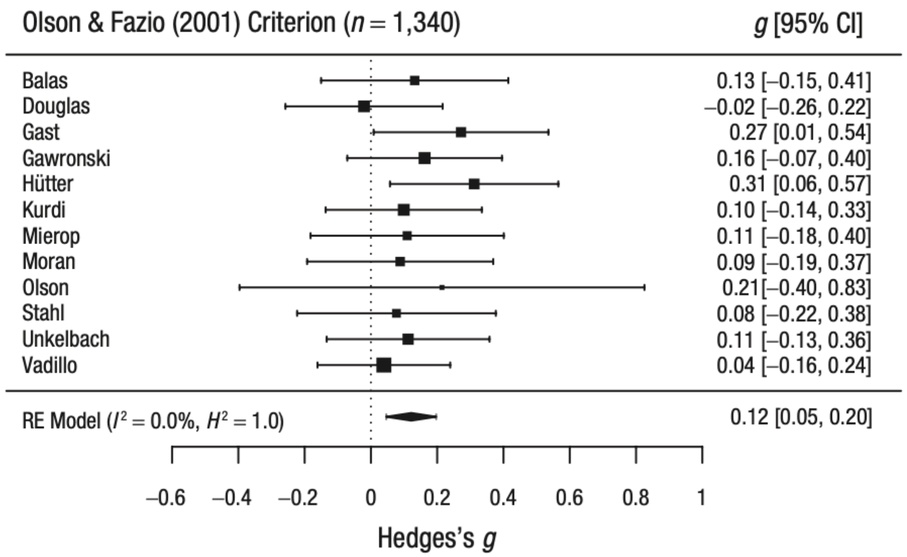
Imagine I made a machine that goes “BING” when it detects that Rummy has had an accident

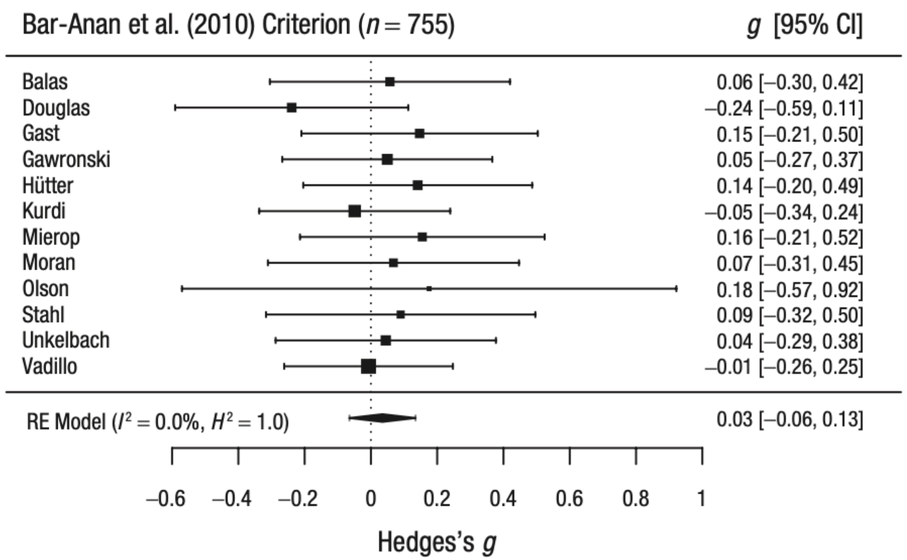
If I do not here a “BING”, should we be certain that I will not find he has 💩 without telling us?

“A t test for independent samples on object-location memory showed no significant effect of priming, t(99) =.50, p =.616, d =.12.”

“The mean proportions of perseverative responses to these tones were very similar between participant groups … with no significant group difference, t(30) = .66, p = .51, d = .23.”

Errors when comparing *p* values





Can people learn attitudes from stimulus pairings they were not aware of?

Moran, Hughes, Hussey, et al. (2021)

Does the exclusion method change the results?

Moran, Hughes, Hussey, et al. (2021)

Does the exclusion method change the results?

Is the effect size moderated by the exclusion method?

Moran, Hughes, Hussey, et al. (2021)

“The difference between
significant and non-significant
is not itself significant”

German & Stern (2006)

Der Unterschied zwischen „statistisch signifikant“ und „statistisch nicht signifikant“ ist selbst nicht statistisch signifikant.

If you want to know whether two things differ, you must directly compare them

E.g., calculate one *p* value for the differences, don’t infer from two *p* values

This error is very prevalent: 50% of neuroscience papers do it wrong! Nieuwenhuis et al. (2011)

Readings

- Aczel et al. (2018)
  - Skim to understand how not to interpret non-significant *p* values
- Gelman & Stern - 2006
  - Skim to understand how not to interpret the pairs of significant and non-significant *p* values
- Nieuwenhuis et al. (2011)
  - Skim to understand the prevalence of the German error: 50%!

- All will be available on Ilias

Assignment

Complete the quiz that uses data from Aczel et al. (2018)

- Will be available on Ilias

Fill in the collaborative Google Sheet with examples from your articles

- The articles assigned to you for StatCheck
- Full instructions will be on Ilias

Questions?