

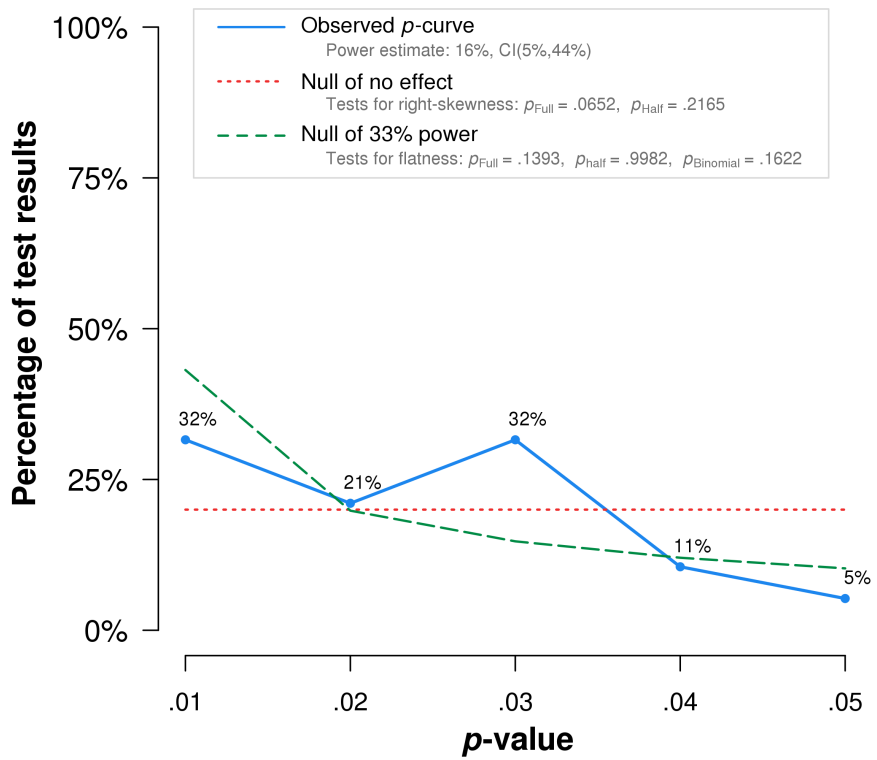
Supplementary Online Material for:

Grebe, N.M., et al.

Testosterone, Cortisol, and Status-Striving Personality Features: A Review and Empirical Evaluation of the Dual Hormone Hypothesis

S1: Full output from p-curve app (www.p-curve.com) of the 19 statistically significant, independent, and directionally consistent Dual Hormone effects. See Simonsohn et al. (2014) for further details.

P-CURVE RESULTS - App 4.06
App's Last Update: 2017 11 30



Note: The observed *p*-curve includes 19 statistically significant ($p < .05$) results, of which 11 are $p < .025$. There were no non-significant results entered.

The image above is in high resolution (400 dpi), you can save it and use in peer-reviewed publications. Below we report the table previous versions of the app embedded in the image above; it includes more details than those reported within the new figure's legend.

	Binomial Test (Share of results $p < .025$)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's < .05)	Half p-curve (p's < .025)
1) Studies contain evidential value. (Right skew)	$p = .3238$	$Z = -1.51, p = .0652$	$Z = -0.78, p = .2165$
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	$p = .1622$	$Z = -1.08, p = .1393$	$Z = 2.91, p = .9982$
Statistical Power			
Power of tests included in p-curve (correcting for selective reporting)	Estimate: 16% 90% Confidence interval: (5% , 44%)		

Interpretation:

P-Curve analysis combines the half and full p-curve to make inferences about evidential value. In particular, if the half p-curve test is right-skewed with $p < .05$ or both the half and full test are right-skewed with $p < .1$, then p-curve analysis indicates the presence of evidential value. This combination test, introduced in Simonsohn, Simmons and Nelson (2015 [.pdf](#)) 'Better P-Curves' paper, is much more robust to ambitious p-hacking than the simple full p-curve test is.

Here neither condition is met; hence p-curve does not indicate evidential value.

Similarly, p-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$. Here neither condition is met; so p-curve does not indicate evidential value is inadequate nor absent.

As with all p-values, these cutoffs are just benchmarks; the lower the p-values are, the less consistent the data are with the respective null hypotheses. A $p = .049$ is essentially the same as a $p = .051$, while a $p = .0001$ is much more compelling than either.

To appreciate the advantage of these combination tests in relation to the previously used full p-curve tests, see [Figure 2](#) and pages 1149-1151 in the 'Better P-Curves' paper ([.pdf](#)) and check out its Supplement 2 ([.pdf](#))

Brief Explanations of Main Results:

1) **Binomial tests** compare the observed proportion of significant results that are $p < .025$ (in this case: 58%) to the expected proportions when there is no effect (50%), and when studies have 1/3 power (71%). This latter number varies (by a few %) as a function of the degrees of freedom of the tests submitted to p -curve.

2) **Continuous tests** are obtained by computing pp -values for each test (probability of at least as extreme a p -value conditional on $p < .05$), and converting them to Z scores ($N(0,1)$). The sum of these Z scores (19 in this case), divided by the square-root of the number of tests included (again: 19 in this case) is the reported Z score in that column (and corresponding p -value). This approach is known as Stouffer's Method. (Prior to App 3.0 we relied on Fisher's method instead. [See "Better P-Curves"](#) paper.)

Note that the binomial and continuous tests are by definition one-sided (e.g., *more* right skewed than flat). We use negative Z values to indicate deviation in the direction of the alternative hypothesis of interest; for example a negative Z value for the Right-Skew test is evidence against the flat null, and thus in favor of Right-Skew.

3) **Statistical power** is obtained by comparing the expected p -curve for each possible value of power between 5% and 99% to the observed p -curve, and selecting the level of power that leads to the expected p -curve that most closely resembles the observed p -curve. (We quantify the similarity with the overall Z arising from aggregating the resulting pp -values via the Stouffer method, pp -values which depend on the assumed level of power). The best fit possible is $Z=0$.

Dropping Highest/Lowest p -values

(Cumulative meta-analysis)

In order to assess the extent to which p -curve's overall results hinge on a few studies, the figure below reports them excluding a progressively larger number of the most extreme p -values originally included in p -curve.

The first column of charts, reports results that first exclude the smallest p -value in p -curve, then the second smallest, and so on. For example, if p -curve contained the following four p -values: $p=.001$, $p=.004$, $p=.01$ and $p=.045$, the 1st marker would report results with all four p -values, the next marker when one excludes $p=.001$, then excluding both $p=.001$ and $p=.004$, and so on.

In the second column one proceeds in opposite order. First excluding $p=.045$, then $p=.045$ and $p=.01$, and so on.

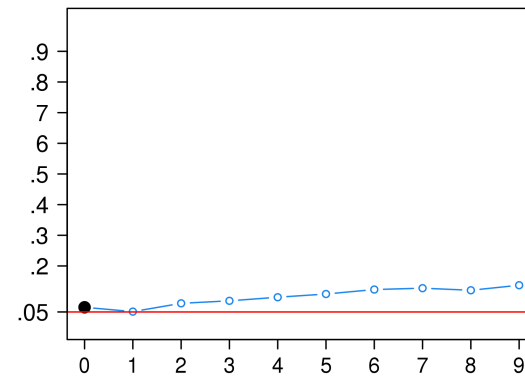
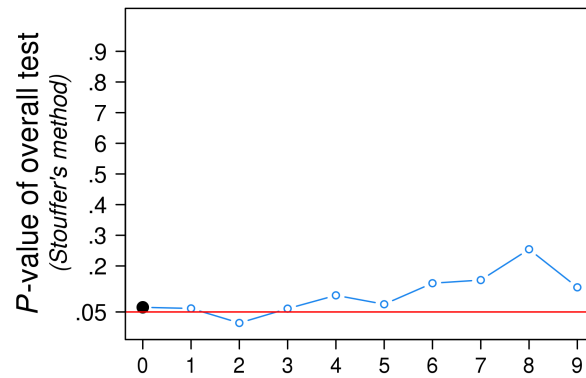
The graph plots what happens until there is only half the p -values left, but in most situations one is only interested in what happens as the single or handful of most extreme p -values are excluded.

We should place more confidence in sets of studies whose overall evidential value survives the exclusion of the most extreme few results.

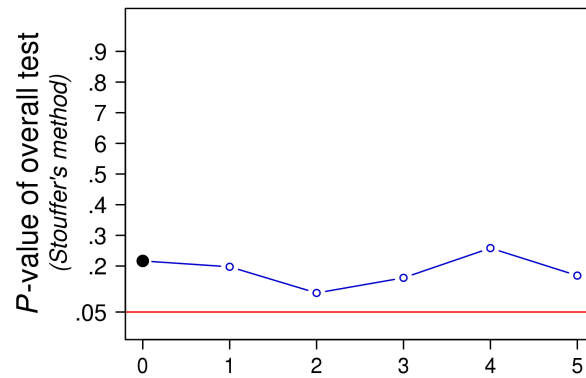
Drop k **lowest** original p -values

Drop k **highest** original p -values

Right skew
(Full p -curve)

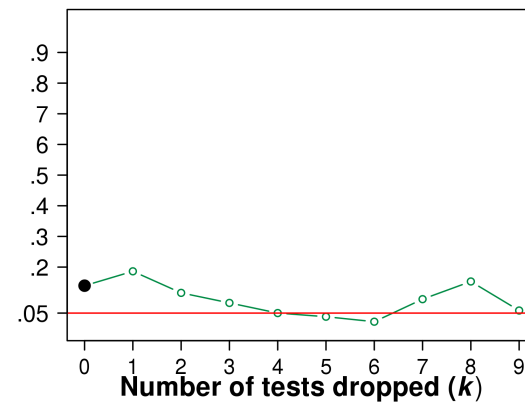
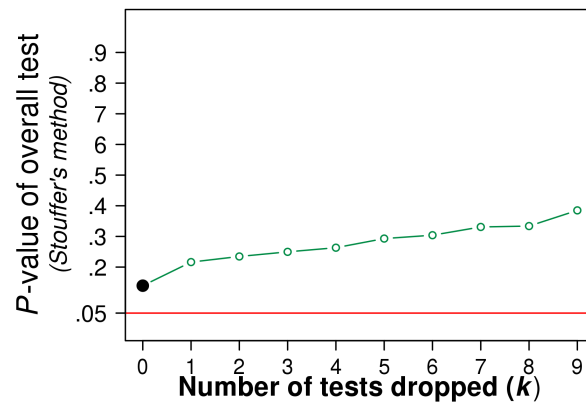


Right skew
(Half p -curve)



Graph not needed
(Half p -curve excludes high p -values)

33% power
(Full p -curve)



● Including all p -values ○ Dropping p -values

Calculations for each test entered into *p*-curve:

Test entered by user	<i>p</i> -value	<i>pp</i> -values				Z Scores			
		Full <i>p</i> -curve		Half <i>p</i> -curve		Full <i>p</i> -curve		Half <i>p</i> -curve	
		Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%
t(67)=2.27	.02643	.52860	.27147	NA	NA	0.07	-0.61	NA	NA
t(112)=2.11	.03708	.74170	.13426	NA	NA	0.65	-1.11	NA	NA
F(1,70)=6.62	.01221	.24412	.51729	.48823	.77279	-0.69	0.04	-0.03	0.75
t(114)=3.17	.00196	.03916	.83127	.07832	.92084	-1.76	0.96	-1.42	1.41
t(451)=2.55	.01110	.22204	.53342	.44408	.78194	-0.77	0.08	-0.14	0.78
F(1,82)=4.96	.02868	.57362	.23936	NA	NA	0.19	-0.71	NA	NA
t(60)=2.028	.04701	.94015	.02926	NA	NA	1.56	-1.89	NA	NA
t(156)=2.67	.00839	.16777	.60271	.33554	.81387	-0.96	0.26	-0.42	0.89
t(106)=2.36	.02010	.40210	.36552	.80419	.70221	-0.25	-0.34	0.86	0.53
t(70)=2.27	.02629	.52579	.27322	NA	NA	0.06	-0.60	NA	NA
t(147)=2.79	.00597	.11940	.67221	.23880	.84639	-1.18	0.45	-0.71	1.02
t(37)=2.26	.02980	.59594	.22936	NA	NA	0.24	-0.74	NA	NA
t(35)=2.34	.02511	.50228	.29610	NA	NA	0.01	-0.54	NA	NA
t(91)=2.37	.01990	.39800	.36972	.79600	.70391	-0.26	-0.33	0.83	0.54
t(39)=2.744	.00913	.18251	.59788	.36502	.80945	-0.91	0.25	-0.35	0.88
t(69)=2.90	.00500	.10003	.70989	.20006	.86343	-1.28	0.55	-0.84	1.10
t(102)=2.17	.03233	.64656	.19134	NA	NA	0.38	-0.87	NA	NA
t(18)=3.19	.00507	.10146	.73779	.20291	.87351	-1.27	0.64	-0.83	1.14
F(1,311)=5.77	.01689	.33776	.41754	.67553	.72763	-0.42	-0.21	0.46	0.61
SUM of Z-Scores in column, dividing by sqrt(N of tests) Z Scores reported under <i>p</i> -curve figure above----->						-1.51	-1.08	-0.78	2.91

Explaining these calculations with an example:

Take the first significant result entered: **t(67)=2.27**. It is associated with a two-sided *p*-value of **0.02643**. *pp*-values are the probability of at least as extreme a significant *p*-value. For right skew we compute these under the null of no effect; because *p*-values would be

distributed uniform between 0 and .05, we simply divide by .05 (multiply by 20) and get the *pp*-value for right skew, that is $0.02643 \times 20 = \mathbf{0.5286}$. One minus that gives us the *pp*-value for left skew (not shown above).

For the *pp*-value under the null that the test is powered to 33% things are a bit more complicated. This explanation will not be quite enough, but: we find the non-centrality parameter for the corresponding distribution and degrees of freedom that gives 33% power. We then evaluate in that non-central distribution the observed test statistic, $t(67)=2.27$, and now divide by 33% rather than 5% because now 1/3 of tests are expected to be $p < .05$ rather than only 5% of them.

More importantly, the interpretation of the *pp*-value for 33% power is as follows. If the underlying effect size were big enough to give the sample of the study obtaining $t(67)=2.27$ 33% power, then with probability **0.27147** we would get a *p*-value of 0.02643 or higher.

For the half *p*-curve we proceed similarly. First, for right skew we divide by .025 (multiply by 40). When a *p*-value is $> .025$ it is not included in half *p*-curve, we see "NA" in the table above. For 33% power, in turn, we use the same non-centrality parameter but this time we divide by the share of *p*-values expected to be $p < .025$ when power is 33%.

The last four columns report the Z-Scores associated with those *pp*-values. So for the full *p*-curve right-skew *pp*-value we had **pp=0.5286**. Evaluating the standard normal distribution in that percentile gives us the reported **Z=0.07**.

Diagnostic plot for power estimation

This figure plots how consistent the observed *p*-curve is with each possible value of power between 5% and 99%.

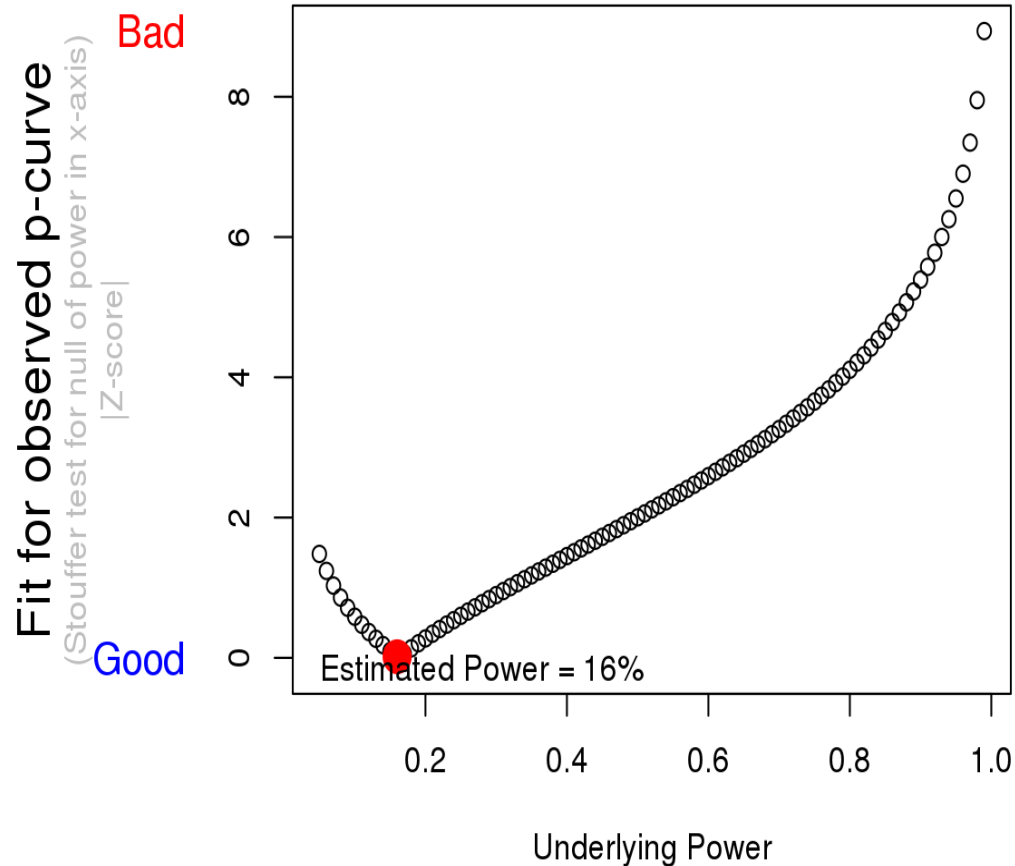
To create the figure we compute *pp*-values for the null that all studies are powered with a given level of power and combine those *pp*-values using Stouffer's method. The best fitting level of power will lead to an overall Stouffer $Z=0$, $p=.5$.

This approach is different from the one used with App 3.0 where instead the Kolmogorov-Smirnov test was run on the resulting distribution of *pp*-values and the uniform. The results with both methods are very similar. The main advantage of the KS test approach is that it reports absolute fit between expected and observed *p*-curve. The main advantage of the Stouffer method is that it is the approach used to compute the confidence interval and is hence more parsimonious.

The table with results at the top of this page reports **16%** as the estimate of power. This means that if all studies in the set were truly powered to 16%, half the time we would see a flatter *p*-curve than the one we see, and half the time we would see a more right-skewed one. So 16% is our best guess.

Estimating underlying statistical power

(Plot should be V shaped, or a smooth line to 99%; else don't trust estimate)



Confidence interval for power

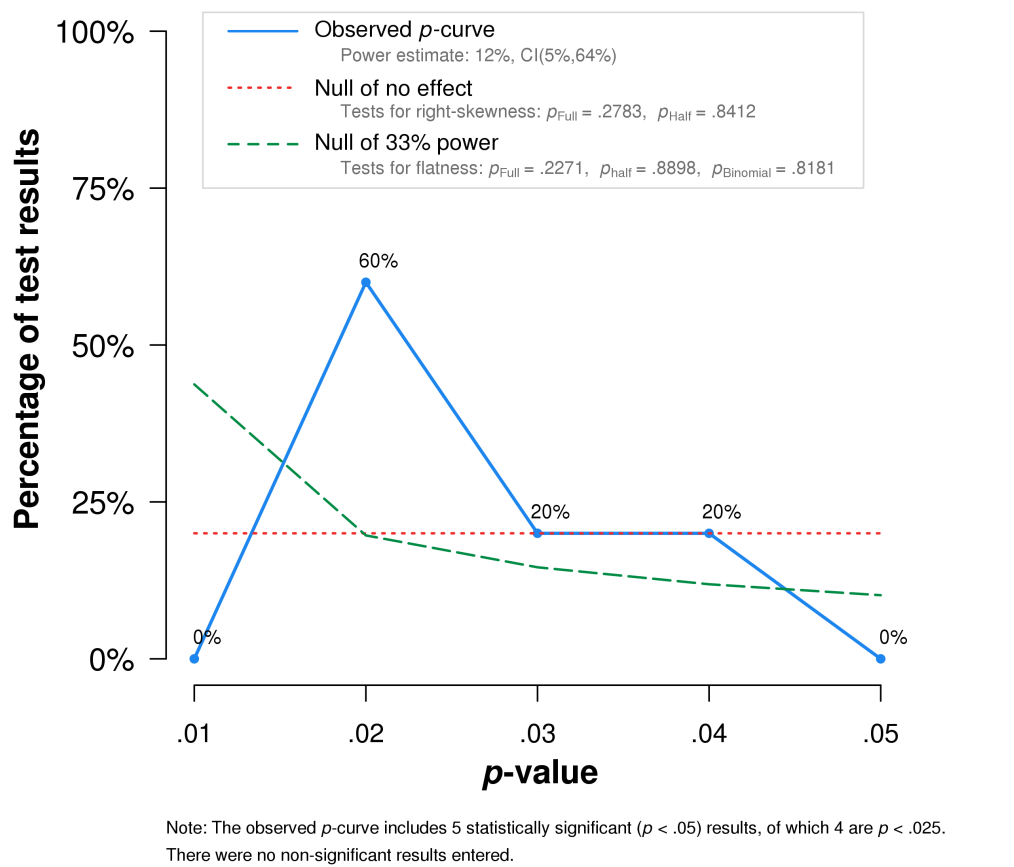
To build the confidence interval for power we proceed as we do to obtain the estimate of power, but rather than finding the underlying statistical power that leads to an overall Stouffer test combining the resulting pp -values of $p=.5$, we find the level of power that gives $p=.05$ and $p=.95$.

For example, above we saw that the lower end of the confidence interval for power was **5%**. This means that if we assume that's the level of power we would observe a p-curve this right-skewed, or more right-skewed, as indexed by the Stouffer combination of the resulting pp-values, only 5% of the time. The other end of the confidence interval (**44%**), in turn, means that if power were that high, we would see as flat a p-curve, or flatter, 95% of the time. Note that this is a 90% confidence interval (for a 95% one, we would look for levels of power leading to overall p-values of 2.5% and 97.5% respectively). We use 90% to make it consistent with the one-sided test against the 33% power null. If p-curve is significantly flatter than expected with 33% power, then the (90%) confidence interval for power will not include 33% power.

Thank you for using the *p*-curve app.

S2: Full output from p-curve app (www.p-curve.com) of the 5 statistically significant, independent, and directionally consistent Dual Hormone effects that employ self-report measures.

P-CURVE RESULTS - App 4.06
App's Last Update: 2017 11 30



The image above is in high resolution (400 dpi), you can save it and use in peer-reviewed publications. Below we report the table previous versions of the app embedded in the image above; it includes more details than those reported within the new figure's legend.

	Binomial Test (Share of results $p < .025$)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p 's $< .05$)	Half p-curve (p 's $< .025$)
1) Studies contain evidential value. (Right skew)	$p = .1875$	$Z = -0.59, p = .2783$	$Z = 1, p = .8412$
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	$p = .8181$	$Z = -0.75, p = .2271$	$Z = 1.23, p = .8898$
	Statistical Power		
Power of tests included in p-curve (correcting for selective reporting)		Estimate: 12% 90% Confidence interval: (5% , 64%)	

Interpretation:

P-Curve analysis combines the half and full *p*-curve to make inferences about evidential value. In particular, if the half *p*-curve test is right-skewed with $p < .05$ or both the half and full test are right-skewed with $p < .1$, then *p*-curve analysis indicates the presence of evidential value. This combination test, introduced in Simonsohn, Simmons and Nelson (2015 [.pdf](#)) 'Better P-Curves' paper, is much more robust to ambitious *p*-hacking than the simple full *p*-curve test is.

Here neither condition is met; hence *p*-curve does not indicate evidential value.

Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$. Here neither condition is met; so *p*-curve does not indicate evidential value is inadequate nor absent.

As with all *p*-values, these cutoffs are just benchmarks; the lower the *p*-values are, the less consistent the data are with the respective null hypotheses. A $p = .049$ is essentially the same as a $p = .051$, while a $p = .0001$ is much more compelling than either.

To appreciate the advantage of these combination tests in relation to the previously used full *p*-curve tests, see [Figure 2](#) and pages 1149-1151 in the 'Better P-Curves' paper ([.pdf](#)) and check out its Supplement 2 ([.pdf](#))

Brief Explanations of Main Results:

1) **Binomial tests** compare the observed proportion of significant results that are $p < .025$ (in this case: 80%) to the expected proportions when there is no effect (50%), and when studies have 1/3 power (71%). This latter number varies (by a few %) as a function of the degrees of freedom of the tests submitted to p -curve.

2) **Continuous tests** are obtained by computing pp -values for each test (probability of at least as extreme a p -value conditional on $p < .05$), and converting them to Z scores ($N(0,1)$). The sum of these Z scores (5 in this case), divided by the square-root of the number of tests included (again: 5 in this case) is the reported Z score in that column (and corresponding p -value). This approach is known as Stouffer's Method. (Prior to App 3.0 we relied on Fisher's method instead. [See "Better P-Curves"](#) paper.)

Note that the binomial and continuous tests are by definition one-sided (e.g., *more* right skewed than flat). We use negative Z values to indicate deviation in the direction of the alternative hypothesis of interest; for example a negative Z value for the Right-Skew test is evidence against the flat null, and thus in favor of Right-Skew.

3) **Statistical power** is obtained by comparing the expected p -curve for each possible value of power between 5% and 99% to the observed p -curve, and selecting the level of power that leads to the expected p -curve that most closely resembles the observed p -curve. (We quantify the similarity with the overall Z arising from aggregating the resulting pp -values via the Stouffer method, pp -values which depend on the assumed level of power). The best fit possible is $Z=0$.

Dropping Highest/Lowest p -values

(Cumulative meta-analysis)

In order to assess the extent to which p -curve's overall results hinge on a few studies, the figure below reports them excluding a progressively larger number of the most extreme p -values originally included in p -curve.

The first column of charts, reports results that first exclude the smallest p -value in p -curve, then the second smallest, and so on. For example, if p -curve contained the following four p -values: $p=.001$, $p=.004$, $p=.01$ and $p=.045$, the 1st marker would report results with all four p -values, the next marker when one excludes $p=.001$, then excluding both $p=.001$ and $p=.004$, and so on.

In the second column one proceeds in opposite order. First excluding $p=.045$, then $p=.045$ and $p=.01$, and so on.

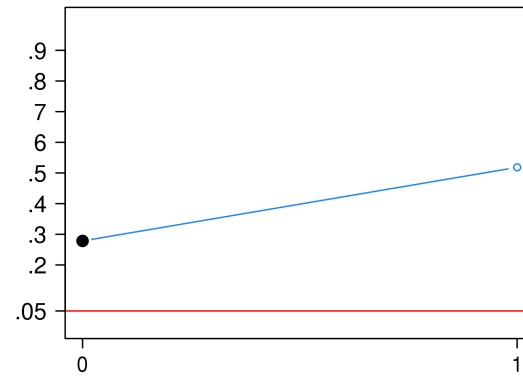
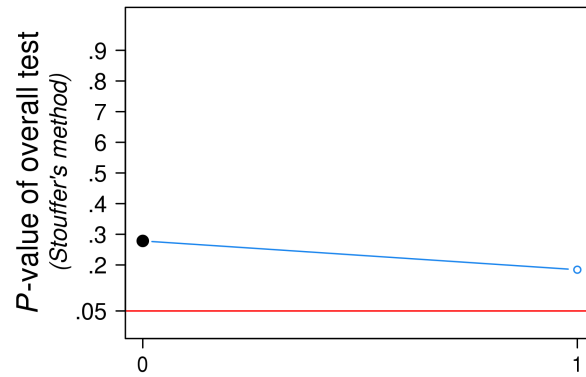
The graph plots what happens until there is only half the p -values left, but in most situations one is only interested in what happens as the single or handful of most extreme p -values are excluded.

We should place more confidence in sets of studies whose overall evidential value survives the exclusion of the most extreme few results.

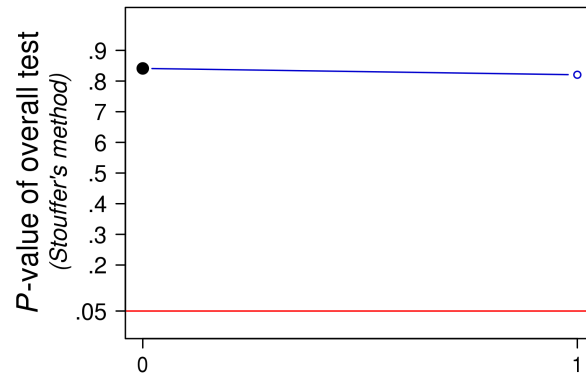
Drop k **lowest** original p -values

Drop k **highest** original p -values

Right skew
(Full p -curve)

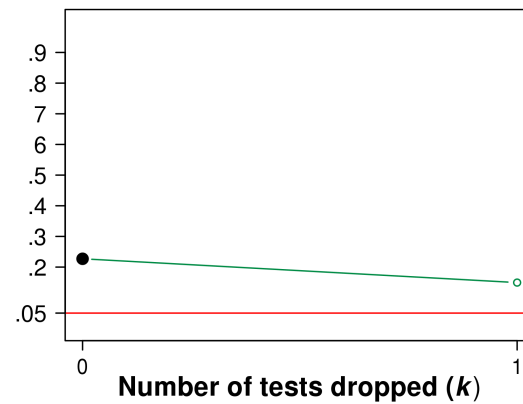
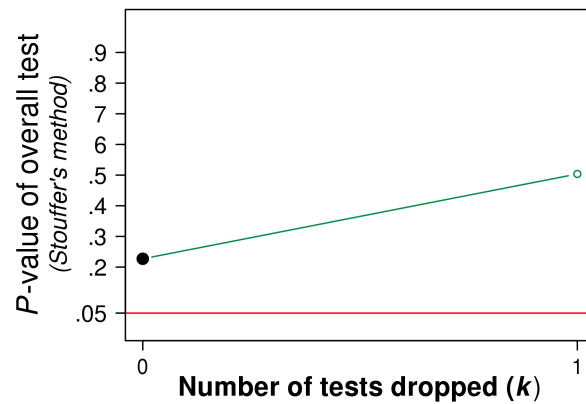


Right skew
(Half p -curve)



Graph not needed
(Half p -curve excludes high p -values)

33% power
(Full p -curve)



● Including all p -values ○ Dropping p -values

Calculations for each test entered into *p*-curve:

Test entered by user	<i>p</i> -value	<i>pp</i> -values				Z Scores			
		Full <i>p</i> -curve		Half <i>p</i> -curve		Full <i>p</i> -curve		Half <i>p</i> -curve	
		Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%
t(451)=2.55	.01110	.22204	.53342	.44408	.78194	-0.77	0.08	-0.14	0.78
t(106)=2.36	.02010	.40210	.36552	.80419	.70221	-0.25	-0.34	0.86	0.53
t(91)=2.37	.01990	.39800	.36972	.79600	.70391	-0.26	-0.33	0.83	0.54
t(102)=2.17	.03233	.64656	.19134	NA	NA	0.38	-0.87	NA	NA
F(1,311)=5.77	.01689	.33776	.41754	.67553	.72763	-0.42	-0.21	0.46	0.61
SUM of Z-Scores in column, dividing by sqrt(N of tests) Z Scores reported under <i>p</i> -curve figure above----->						-0.59	-0.75	1	1.23

Explaining these calculations with an example:

Take the first significant result entered: **t(451)=2.55**. It is associated with a two-sided *p*-value of **0.0111**. *pp*-values are the probability of at least as extreme a significant *p*-value. For right skew we compute these under the null of no effect; because *p*-values would be distributed uniform between 0 and .05, we simply divide by .05 (multiply by 20) and get the *pp*-value for right skew, that is $0.0111 \times 20 = \mathbf{0.22204}$. One minus that gives us the *pp*-value for left skew (not shown above).

For the *pp*-value under the null that the test is powered to 33% things are a bit more complicated. This explanation will not be quite enough, but: we find the non-centrality parameter for the corresponding distribution and degrees of freedom that gives 33% power. We then evaluate in that non-central distribution the observed test statistic, t(451)=2.55, and now divide by 33% rather than 5% because now 1/3 of tests are expected to be $p < .05$ rather than only 5% of them.

More importantly, the interpretation of the *pp*-value for 33% power is as follows. If the underlying effect size were big enough to give the sample of the study obtaining t(451)=2.55 33% power, then with probability **0.53342** we would get a *p*-value of 0.0111 or higher.

For the half *p*-curve we proceed similarly. First, for right skew we divide by .025 (multiply by 40). When a *p*-value is $> .025$ it is not included in half *p*-curve, we see "NA" in the table above. For 33% power, in turn, we use the same non-centrality parameter but this time we divide by the share of *p*-values expected to be $p < .025$ when power is 33%.

The last four columns report the Z-Scores associated with those *pp*-values. So for the full *p*-curve right-skew *pp*-value we had **pp=0.22204**. Evaluating the standard normal distribution in that percentile gives us the reported **Z=-0.77**.

Diagnostic plot for power estimation

This figure plots how consistent the observed p -curve is with each possible value of power between 5% and 99%.

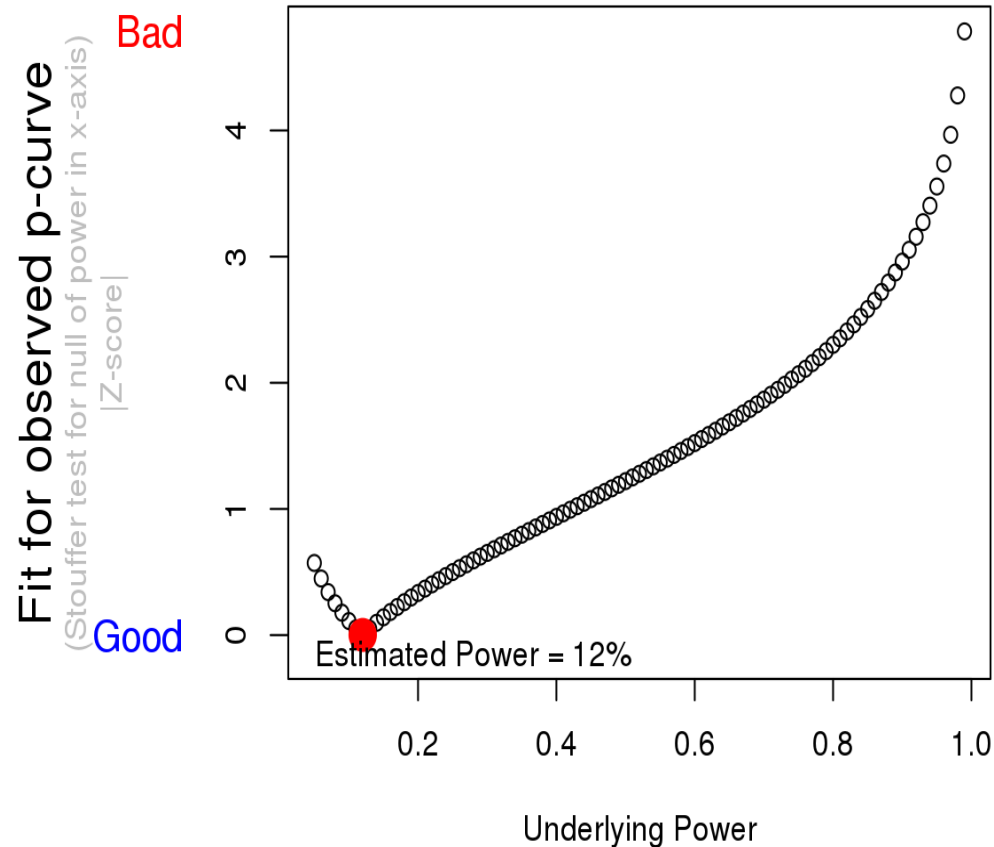
To create the figure we compute pp -values for the null that all studies are powered with a given level of power and combine those pp -values using Stouffer's method. The best fitting level of power will lead to an overall Stouffer $Z=0$, $p=.5$.

This approach is different from the one used with App 3.0 where instead the Kolmogorov-Smirnov test was run on the resulting distribution of pp -values and the uniform. The results with both methods are very similar. The main advantage of the KS test approach is that it reports absolute fit between expected and observed p -curve. The main advantage of the Stouffer method is that it is the approach used to compute the confidence interval and is hence more parsimonious.

The table with results at the top of this page reports **12%** as the estimate of power. This means that if all studies in the set were truly powered to 12%, half the time we would see a flatter p -curve than the one we see, and half the time we would see a more right-skewed one. So 12% is our best guess.

Estimating underlying statistical power

(Plot should be V shaped, or a smooth line to 99%; else don't trust estimate)



Confidence interval for power

To build the confidence interval for power we proceed as we do to obtain the estimate of power, but rather than finding the underlying statistical power that leads to an overall Stouffer test combining the resulting p -values of $p=.5$, we find the level of power that gives $p=.05$ and $p=.95$.

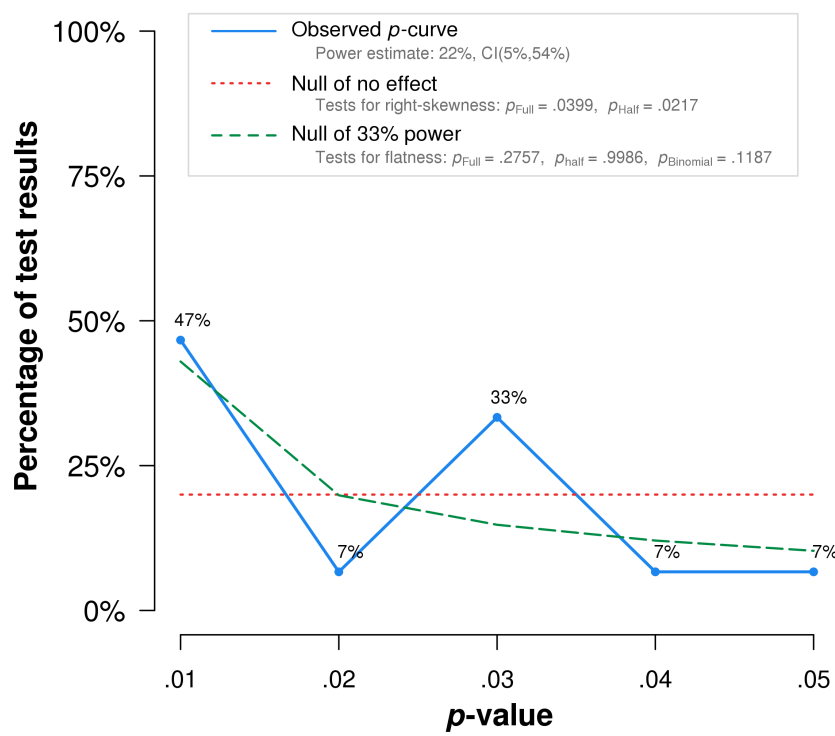
For example, above we saw that the lower end of the confidence interval for power was **5%**. This means that if we assume that's the

level of power we would observe a p-curve this right-skewed, or more right-skewed, as indexed by the Stouffer combination of the resulting p-values, only 5% of the time. The other end of the confidence interval (**64%**), in turn, means that if power were that high, we would see as flat a p-curve, or flatter, 95% of the time. Note that this is a 90% confidence interval (for a 95% one, we would look for levels of power leading to overall p-values of 2.5% and 97.5% respectively). We use 90% to make it consistent with the one-sided test against the 33% power null. If p-curve is significantly flatter than expected with 33% power, then the (90%) confidence interval for power will not include 33% power.

Thank you for using the *p*-curve app.

S3: Full output from p-curve app (www.p-curve.com) of the 15 statistically significant, independent, and directionally consistent Dual Hormone effects that employ behavioral measures.

P-CURVE RESULTS - App 4.06
App's Last Update: 2017 11 30



Note: The observed *p*-curve includes 15 statistically significant ($p < .05$) results, of which 8 are $p < .025$. There were no non-significant results entered.

The image above is in high resolution (400 dpi), you can save it and use in peer-reviewed publications. Below we report the table previous versions of the app embedded in the image above; it includes more details than those reported within the new figure's legend.

	Binomial Test (Share of results $p < .025$)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's < .05)	Half p-curve (p's < .025)
1) Studies contain evidential value. (Right skew)	$p = .5$	$Z = -1.75, p = .0399$	$Z = -2.02, p = .0217$
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	$p = .1187$	$Z = -0.6, p = .2757$	$Z = 2.98, p = .9986$
	Statistical Power		
Power of tests included in p-curve (correcting for selective reporting)		Estimate: 22% 90% Confidence interval: (5% , 54%)	

Interpretation:

P-Curve analysis combines the half and full p-curve to make inferences about evidential value. In particular, if the half p-curve test is right-skewed with $p < .05$ or both the half and full test are right-skewed with $p < .1$, then p-curve analysis indicates the presence of evidential value. This combination test, introduced in Simonsohn, Simmons and Nelson (2015 [.pdf](#)) 'Better P-Curves' paper, is much more robust to ambitious p-hacking than the simple full p-curve test is.

Here both conditions are met, indicating evidential value.

Similarly, p-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$. Here neither condition is met; so p-curve does not indicate evidential value is inadequate nor absent.

As with all p-values, these cutoffs are just benchmarks; the lower the p-values are, the less consistent the data are with the respective null hypotheses. A $p = .049$ is essentially the same as a $p = .051$, while a $p = .0001$ is much more compelling than either.

To appreciate the advantage of these combination tests in relation to the previously used full p-curve tests, see [Figure 2](#) and pages 1149-1151 in the 'Better P-Curves' paper ([.pdf](#)) and check out its Supplement 2 ([.pdf](#))

Brief Explanations of Main Results:

1) **Binomial tests** compare the observed proportion of significant results that are $p < .025$ (in this case: 53%) to the expected proportions when there is no effect (50%), and when studies have 1/3 power (71%). This latter number varies (by a few %) as a function of the degrees of freedom of the tests submitted to p -curve.

2) **Continuous tests** are obtained by computing pp -values for each test (probability of at least as extreme a p -value conditional on $p < .05$), and converting them to Z scores ($N(0,1)$). The sum of these Z scores (15 in this case), divided by the square-root of the number of tests included (again: 15 in this case) is the reported Z score in that column (and corresponding p -value). This approach is known as Stouffer's Method. (Prior to App 3.0 we relied on Fisher's method instead. [See "Better P-Curves"](#) paper.)

Note that the binomial and continuous tests are by definition one-sided (e.g., *more* right skewed than flat). We use negative Z values to indicate deviation in the direction of the alternative hypothesis of interest; for example a negative Z value for the Right-Skew test is evidence against the flat null, and thus in favor of Right-Skew.

3) **Statistical power** is obtained by comparing the expected p -curve for each possible value of power between 5% and 99% to the observed p -curve, and selecting the level of power that leads to the expected p -curve that most closely resembles the observed p -curve. (We quantify the similarity with the overall Z arising from aggregating the resulting pp -values via the Stouffer method, pp -values which depend on the assumed level of power). The best fit possible is $Z=0$.

Dropping Highest/Lowest p -values

(Cumulative meta-analysis)

In order to assess the extent to which p -curve's overall results hinge on a few studies, the figure below reports them excluding a progressively larger number of the most extreme p -values originally included in p -curve.

The first column of charts, reports results that first exclude the smallest p -value in p -curve, then the second smallest, and so on. For example, if p -curve contained the following four p -values: $p=.001$, $p=.004$, $p=.01$ and $p=.045$, the 1st marker would report results with all four p -values, the next marker when one excludes $p=.001$, then excluding both $p=.001$ and $p=.004$, and so on.

In the second column one proceeds in opposite order. First excluding $p=.045$, then $p=.045$ and $p=.01$, and so on.

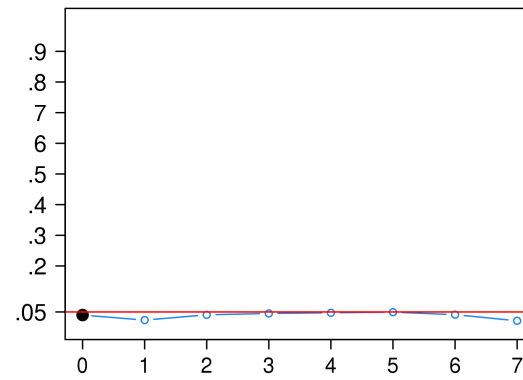
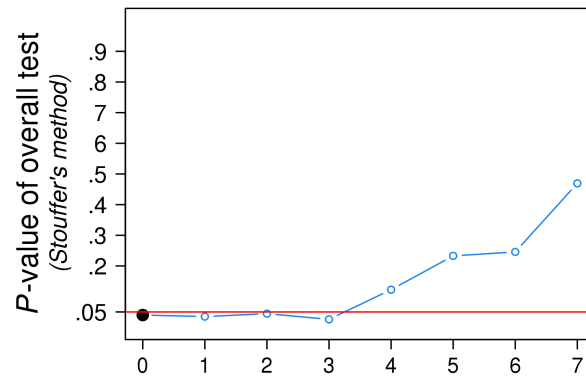
The graph plots what happens until there is only half the p -values left, but in most situations one is only interested in what happens as the single or handful of most extreme p -values are excluded.

We should place more confidence in sets of studies whose overall evidential value survives the exclusion of the most extreme few results.

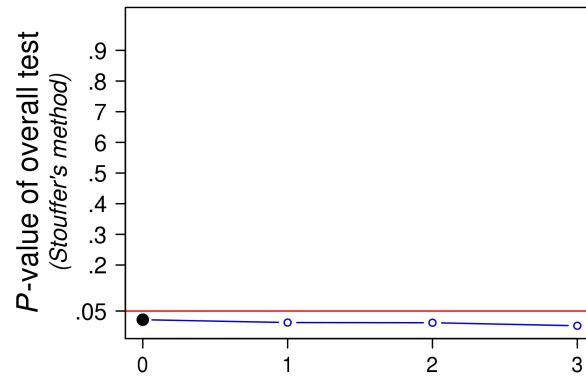
Drop k **lowest** original p -values

Drop k **highest** original p -values

Right skew
(Full p -curve)

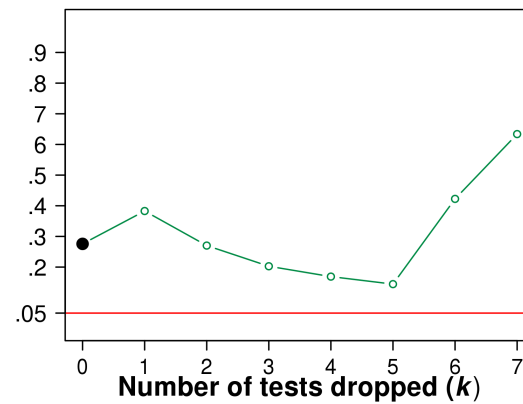
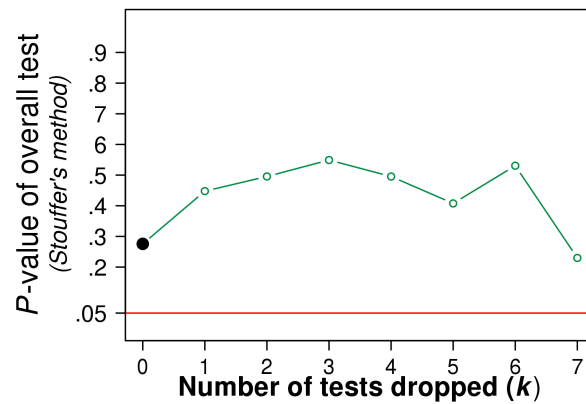


Right skew
(Half p -curve)



Graph not needed
(Half p -curve excludes high p -values)

33% power
(Full p -curve)



● Including all p -values ○ Dropping p -values

Calculations for each test entered into *p*-curve:

Test entered by user	<i>p</i> -value	<i>pp</i> -values				Z Scores			
		Full <i>p</i> -curve		Half <i>p</i> -curve		Full <i>p</i> -curve		Half <i>p</i> -curve	
		Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%
t(67)=2.27	.02643	.52860	.27147	NA	NA	0.07	-0.61	NA	NA
t(112)=2.11	.03708	.74170	.13426	NA	NA	0.65	-1.11	NA	NA
F(1,70)=6.62	.01221	.24412	.51729	.48823	.77279	-0.69	0.04	-0.03	0.75
t(114)=3.17	.00196	.03916	.83127	.07832	.92084	-1.76	0.96	-1.42	1.41
F(1,82)=4.96	.02868	.57362	.23936	NA	NA	0.19	-0.71	NA	NA
t(60)=2.028	.04701	.94015	.02926	NA	NA	1.56	-1.89	NA	NA
t(156)=2.67	.00839	.16777	.60271	.33554	.81387	-0.96	0.26	-0.42	0.89
t(88)=3.02	.00331	.06616	.77078	.13232	.89229	-1.51	0.74	-1.12	1.24
t(70)=2.27	.02629	.52579	.27322	NA	NA	0.06	-0.60	NA	NA
t(147)=2.79	.00597	.11940	.67221	.23880	.84639	-1.18	0.45	-0.71	1.02
t(37)=2.26	.02980	.59594	.22936	NA	NA	0.24	-0.74	NA	NA
t(35)=2.34	.02511	.50228	.29610	NA	NA	0.01	-0.54	NA	NA
t(39)=2.744	.00913	.18251	.59788	.36502	.80945	-0.91	0.25	-0.35	0.88
t(69)=2.90	.00500	.10003	.70989	.20006	.86343	-1.28	0.55	-0.84	1.10
t(18)=3.19	.00507	.10146	.73779	.20291	.87351	-1.27	0.64	-0.83	1.14
SUM of Z-Scores in column, dividing by sqrt(N of tests) Z Scores reported under <i>p</i> -curve figure above----->						-1.75	-0.6	-2.02	2.98

Explaining these calculations with an example:

Take the first significant result entered: **t(67)=2.27**. It is associated with a two-sided *p*-value of **0.02643**. *pp*-values are the probability of at least as extreme a significant *p*-value. For right skew we compute these under the null of no effect; because *p*-values would be distributed uniform between 0 and .05, we simply divide by .05 (multiply by 20) and get the *pp*-value for right skew, that is $0.02643 \times 20 = 0.5286$. One minus that gives us the *pp*-value for left skew (not shown above).

For the *pp*-value under the null that the test is powered to 33% things are a bit more complicated. This explanation will not be quite enough, but: we find the non-centrality parameter for the corresponding distribution and degrees of freedom that gives 33% power. We then evaluate in that non-central distribution the observed test statistic, $t(67)=2.27$, and now divide by 33% rather than 5% because

now 1/3 of tests are expected to be $p < .05$ rather than only 5% of them.

More importantly, the interpretation of the pp -value for 33% power is as follows. If the underlying effect size were big enough to give the sample of the study obtaining $t(67)=2.27$ 33% power, then with probability **0.27147** we would get a p -value of 0.02643 or higher.

For the half p -curve we proceed similarly. First, for right skew we divide by .025 (multiply by 40). When a p -value is $>.025$ it is not included in half p -curve, we see "NA" in the table above. For 33% power, in turn, we use the same non-centrality parameter but this time we divide by the share of p -values expected to be $p < .025$ when power is 33%.

The last four columns report the Z-Scores associated with those pp -values. So for the full p -curve right-skew pp -value we had **$pp=0.5286$** . Evaluating the standard normal distribution in that percentile gives us the reported **$Z=0.07$** .

Diagnostic plot for power estimation

This figure plots how consistent the observed p -curve is with each possible value of power between 5% and 99%.

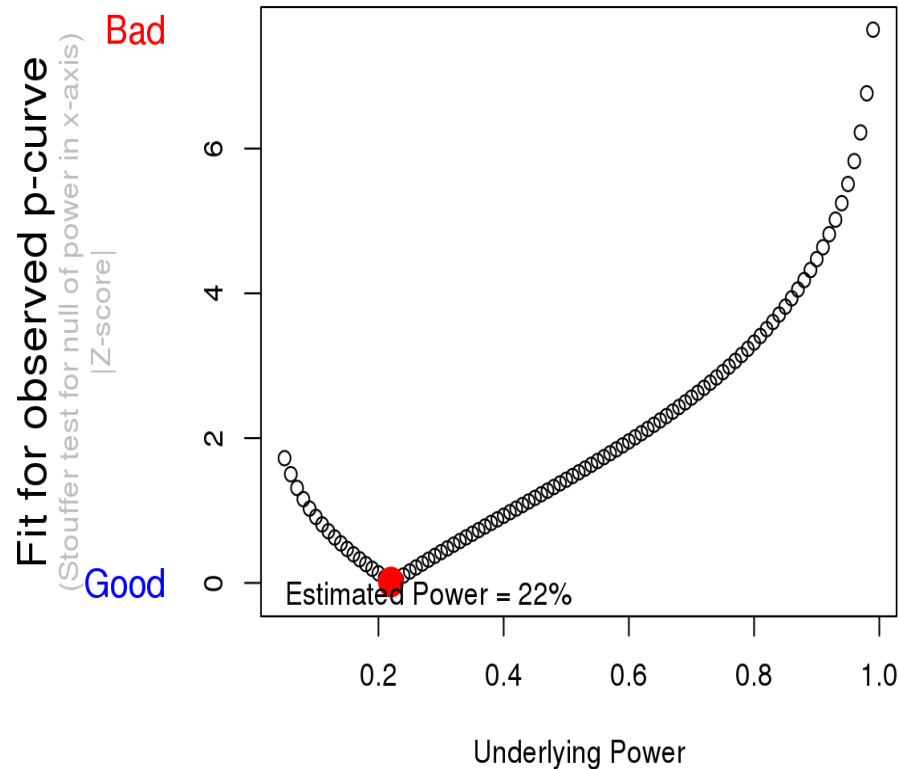
To create the figure we compute pp -values for the null that all studies are powered with a given level of power and combine those pp -values using Stouffer's method. The best fitting level of power will lead to an overall Stouffer $Z=0$, $p=.5$.

This approach is different from the one used with App 3.0 where instead the Kolmogorov-Smirnov test was run on the resulting distribution of pp -values and the uniform. The results with both methods are very similar. The main advantage of the KS test approach is that it reports absolute fit between expected and observed p -curve. The main advantage of the Stouffer method is that it is the approach used to compute the confidence interval and is hence more parsimonious.

The table with results at the top of this page reports **22%** as the estimate of power. This means that if all studies in the set were truly powered to 22%, half the time we would see a flatter p -curve than the one we see, and half the time we would see a more right-skewed one. So 22% is our best guess.

Estimating underlying statistical power

(Plot should be V shaped, or a smooth line to 99%; else don't trust estimate)



Confidence interval for power

To build the confidence interval for power we proceed as we do to obtain the estimate of power, but rather than finding the underlying statistical power that leads to an overall Stouffer test combining the resulting pp -values of $p=.5$, we find the level of power that gives $p=.05$ and $p=.95$.

For example, above we saw that the lower end of the confidence interval for power was **5%**. This means that if we assume that's the level of power we would observe a p-curve this right-skewed, or more right-skewed, as indexed by the Stouffer combination of the resulting pp -values, only 5% of the time. The other end of the confidence interval (**54%**), in turn, means that if power were that high,

we would see as flat a p-curve, or flatter, 95% of the time. Note that this is a 90% confidence interval (for a 95% one, we would look for levels of power leading to overall p-values of 2.5% and 97.5% respectively). We use 90% to make it consistent with the one-sided test against the 33% power null. If p-curve is significantly flatter than expected with 33% power, then the (90%) confidence interval for power will not include 33% power.

Thank you for using the *p*-curve app.

S4: Bayes Factor Analyses

Below we report Bayes factors comparing the null hypothesis of no interaction effect (H_0) with the alternative hypothesis of non-zero interaction (H_1). We use the label BF_{01} to indicate the direction of the comparison. A Bayes factor is the relative likelihood of the data under the two hypotheses; thus, values of BF_{01} larger than 1 indicate evidence in support of H_0 , whereas values of BF_{01} smaller than 1 indicate evidence in support of H_1 . Bayes factors between 1 and 3 (between 0.33 and 1 for H_1) are usually regarded as weak or anecdotal evidence, while values of 10 or more (0.1 or less for H_1) are regarded as strong evidence. Importantly, the Dual Hormone hypothesis predicts negative (attenuating) interactions between T and C levels. Thus, a nonzero interaction effect may still fail to support the dual hormone hypothesis if the direction is positive (potentiating). For each effect, we report both the value of BF_{01} and the direction of the interaction (negative or positive; the latter includes cases in which the effect is approximately zero).

Note that Bayes factors do not directly quantify the odds of H_0 being true (over H_1), but only the *change* in the odds given the data at hand (see Rouder et al., 2018; Wagenmakers et al., 2017). For example, if H_0 and H_1 are initially regarded as equally probable, a Bayes factor $BF_{01} = 5$ in favor of the null means that H_0 should now be regarded as 5 times more probable than H_1 . However, if H_0 is initially assigned a .90 probability (i.e., 9 times more probable than H_1), a Bayes factor $BF_{01} = 5$ in favor of the null means that H_0 should now be regarded as 45 times more probable than H_1 .

For the alternative hypothesis (H_1), we used a default JZS prior with scale factor $r = 0.2$, somewhat lower than the commonly employed $r = 0.354$. This implies a prior distribution of effect sizes in which half of the standardized regression coefficients are within ± 0.2 (see Rouder & Morey, 2012). This choice of prior is sensible in view of the fact that interaction effects in correlational studies tend to be small even in presence of strong interactions (McClelland & Judd, 1993). Robustness checks on the scale factor for the prior were performed by varying r from 0 to 1.5 (shown in the figures). All analyses were performed with JASP 0.8.6 (JASP Team, 2018) using summary statistics from regression models.

JASP Team (2018). *JASP* (Version 0.8.6) [Computer software] <https://jasp-stats.org>.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376-390.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, doi:10.3758/s13423-017-1420-7.

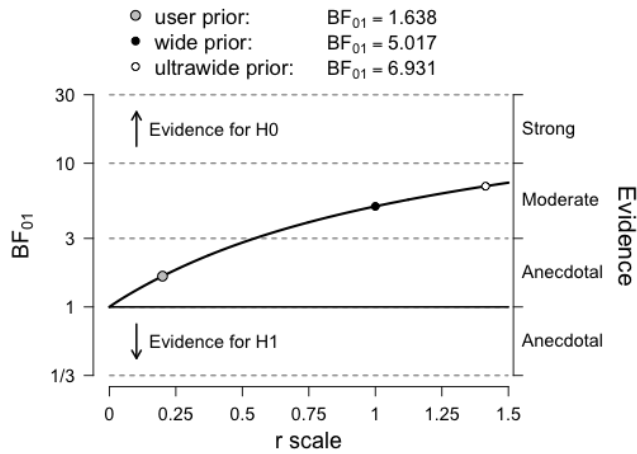
Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47, 877-903.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q.F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J.N., & Morey, R.D. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, doi:10.3758/s13423-017-1343-3.

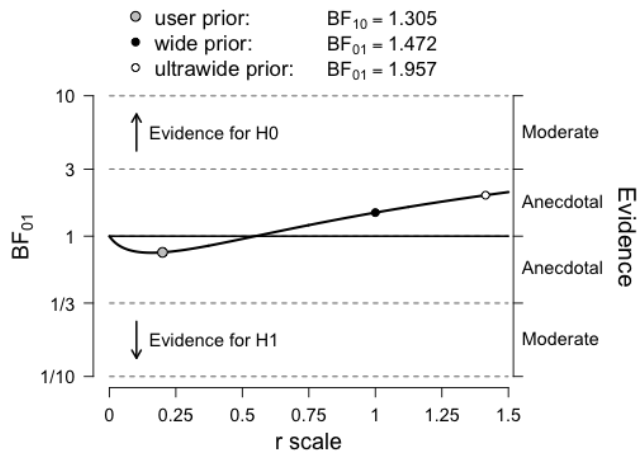
1. Core Traits

1a. Social potency (self-report)

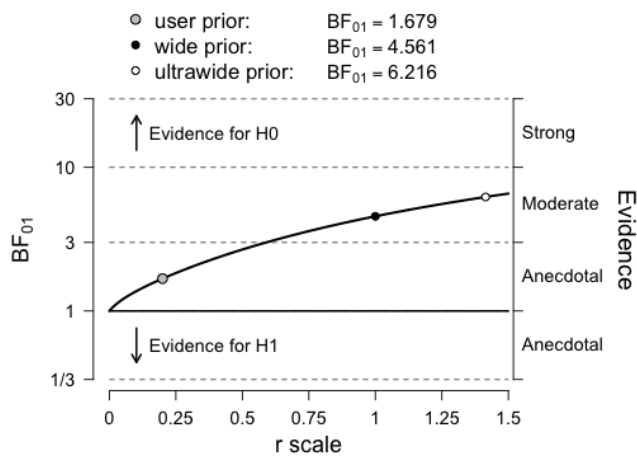
Sample 1, Males: **positive** T×C interaction, $BF_{01} = 1.638$



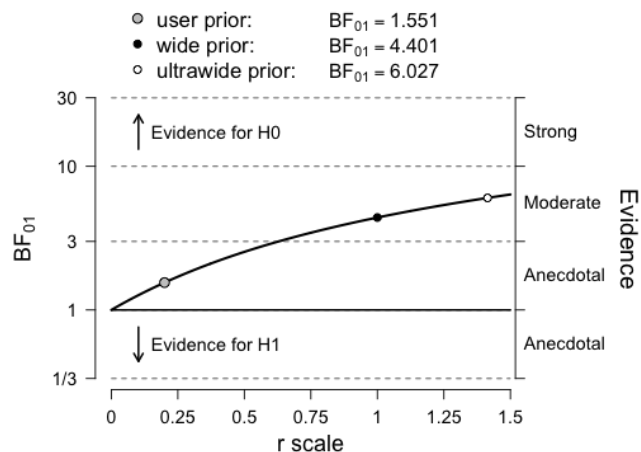
Sample 2, Males: **positive** T×C interaction, $BF_{01} = 0.766$



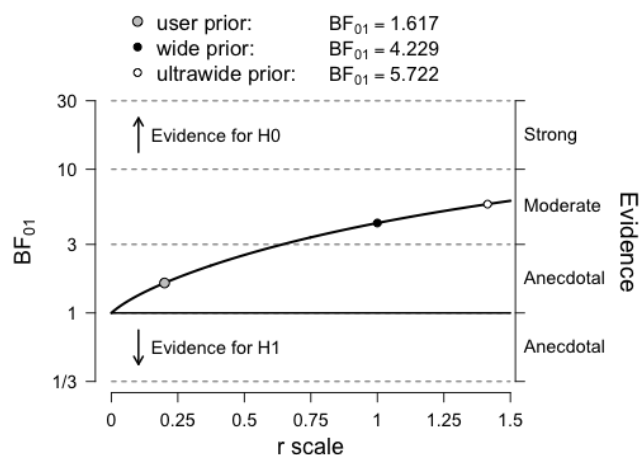
Sample 2, Females: **negative** T×C interaction, $BF_{01} = 1.679$



Sample 3, Males: **positive** T×C interaction, $BF_{01} = 1.551$

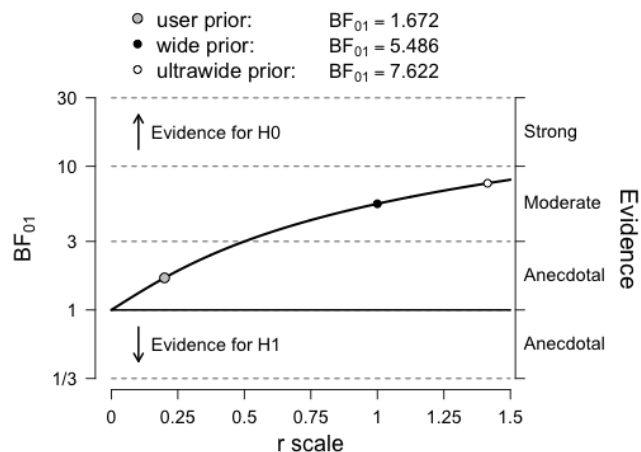


Sample 3, Females: **positive** T×C interaction, $BF_{01} = 1.617$

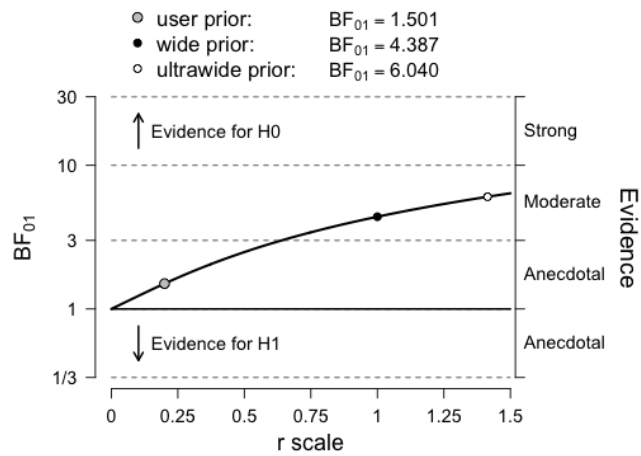


1b. Non-submissiveness (self-report)

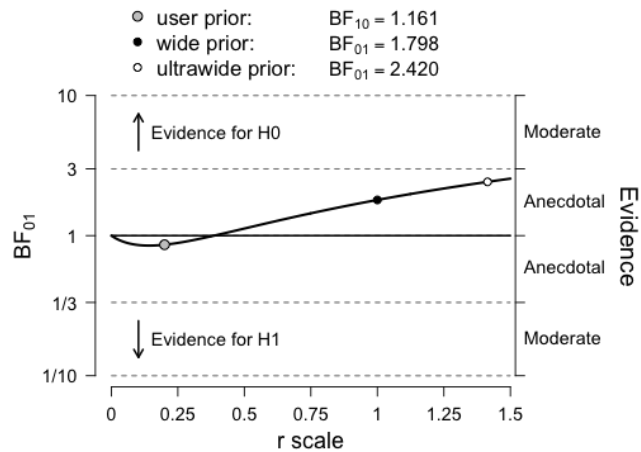
Sample 1, Males: **positive** T×C interaction, $BF_{01} = 1.672$



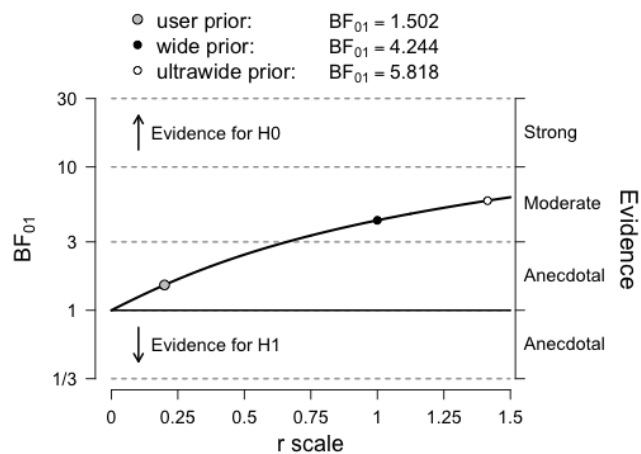
Sample 2, Males: **positive** T×C interaction, $BF_{01} = 1.501$



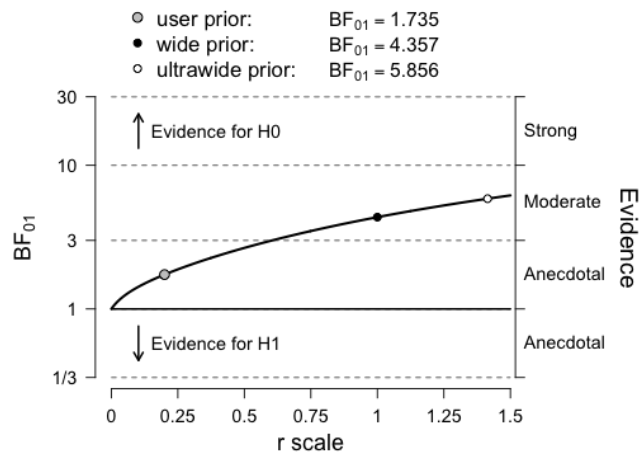
Sample 2, Females: **positive** T×C interaction, $BF_{01} = 0.861$



Sample 3, Males: **positive** T×C interaction, $BF_{01} = 1.502$

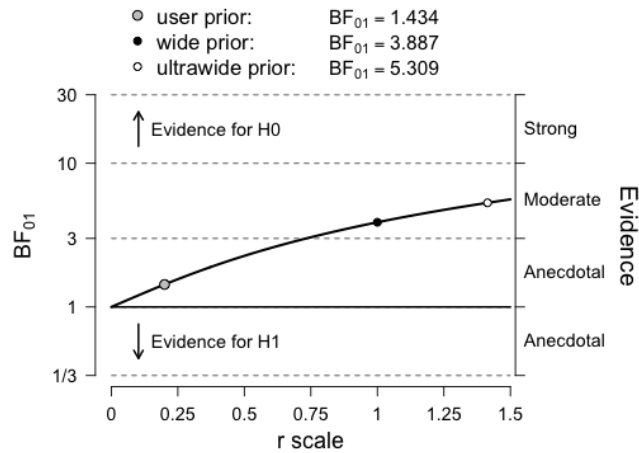


Sample 3, Females: **positive** T×C interaction, $BF_{01} = 1.735$

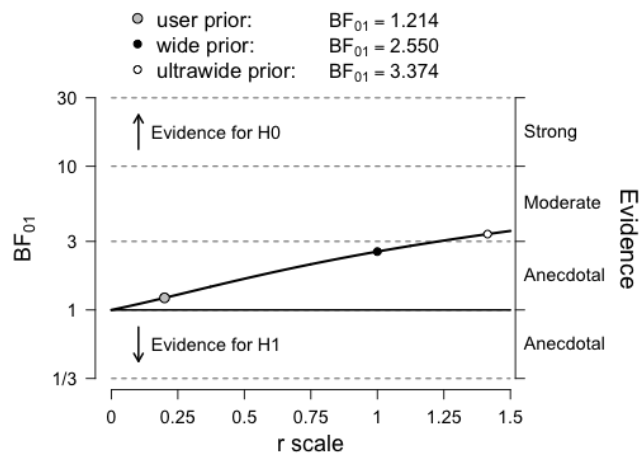


1c. Social potency (partner-report)

Sample 3, Males: **positive** T×C interaction, $BF_{01} = 1.434$

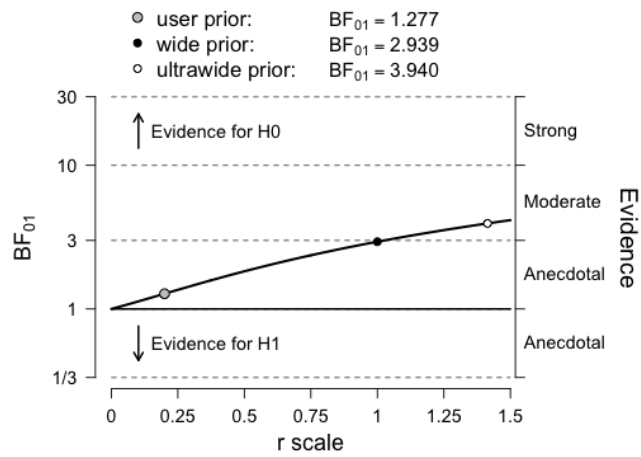


Sample 3, Females: **negative** T×C interaction, $BF_{01} = 1.214$

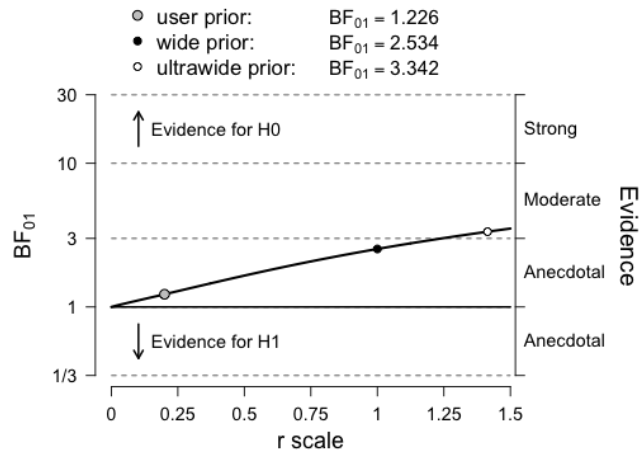


1d. Non-submissiveness (partner-report)

Sample 3, Males: **positive** T×C interaction, $BF_{01} = 1.277$

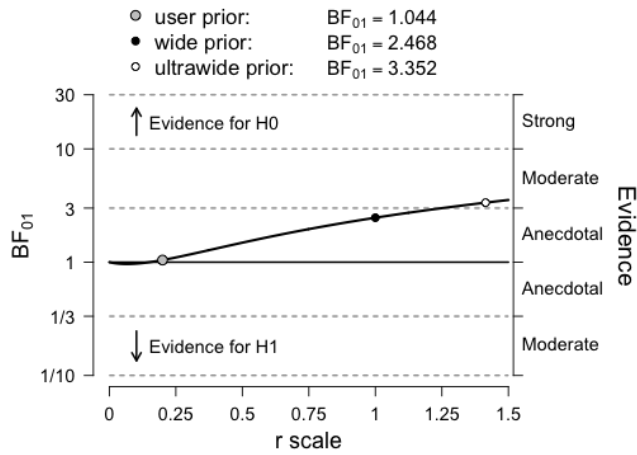


Sample 3, Females: **negative** T×C interaction, $BF_{01} = 1.226$

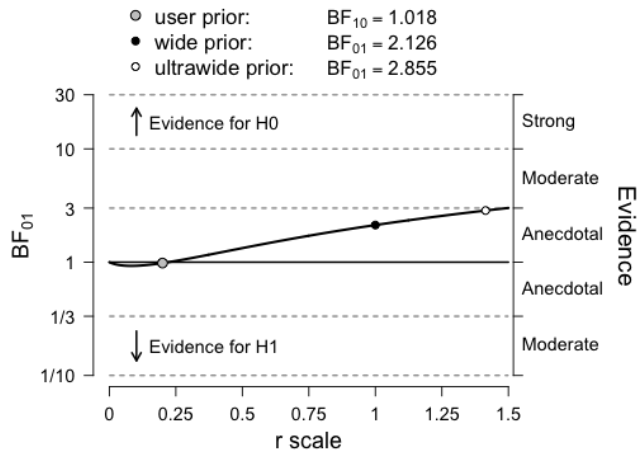


1e. Win intrasexual competition

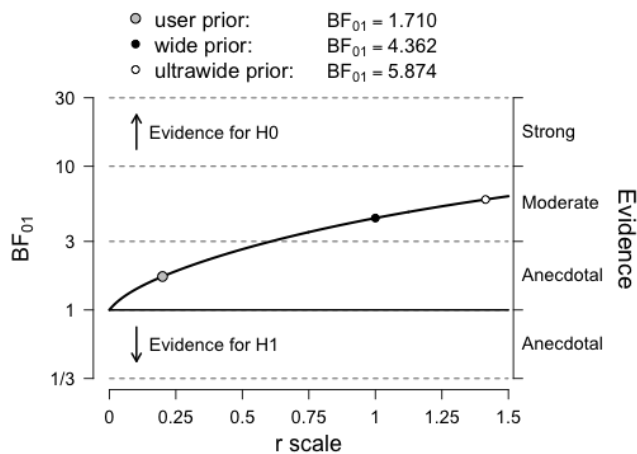
Sample 2, Males: **positive** T×C interaction, $BF_{01} = 1.044$



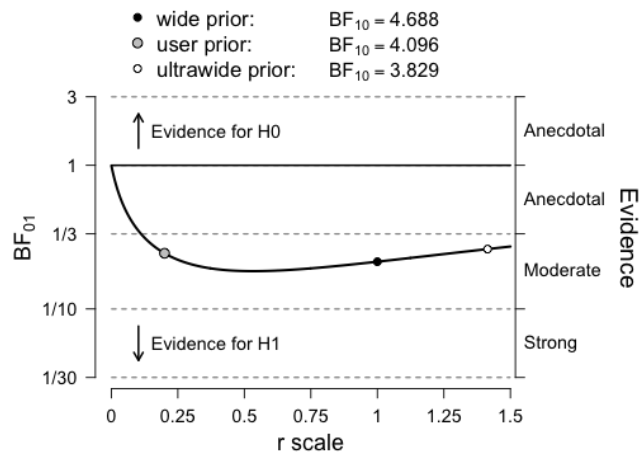
Sample 2, Females: **negative** T×C interaction, $BF_{01} = 0.982$



Sample 3, Males: **negative** T×C interaction, $BF_{01} = 1.710$

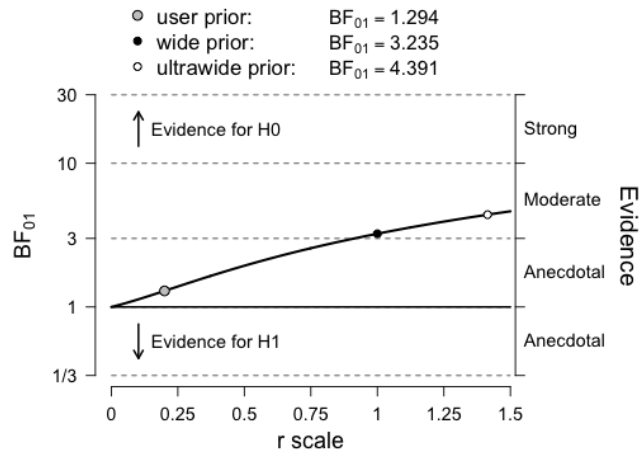


Sample 3, Females: **positive** T×C interaction, $BF_{01} = 0.244$

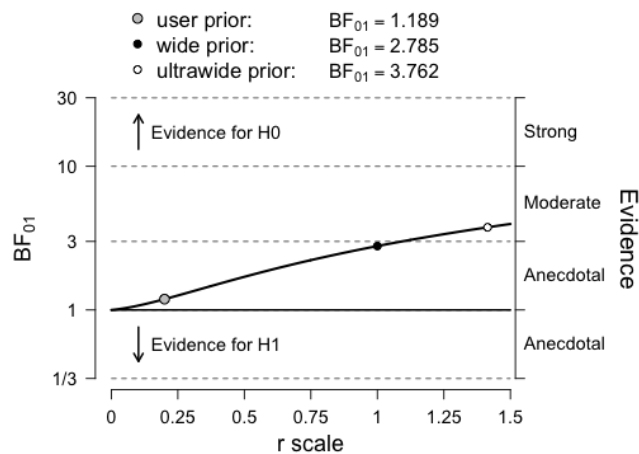


1f. Intrasexual competitiveness (SIC)

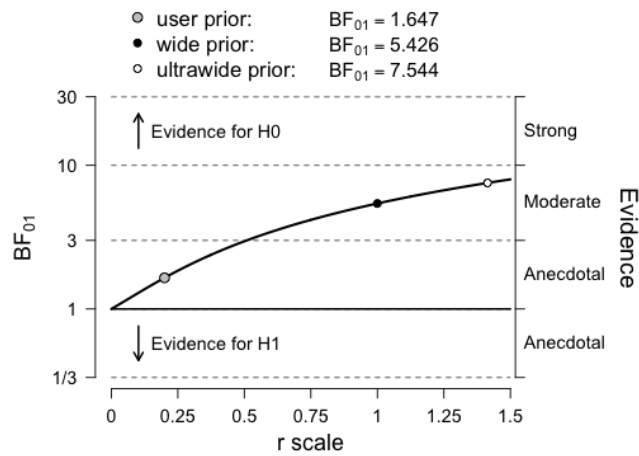
Sample 4, Males: **negative** T×C interaction, $BF_{01} = 1.294$



Sample 4, Females: **negative** T×C interaction, $BF_{01} = 1.189$

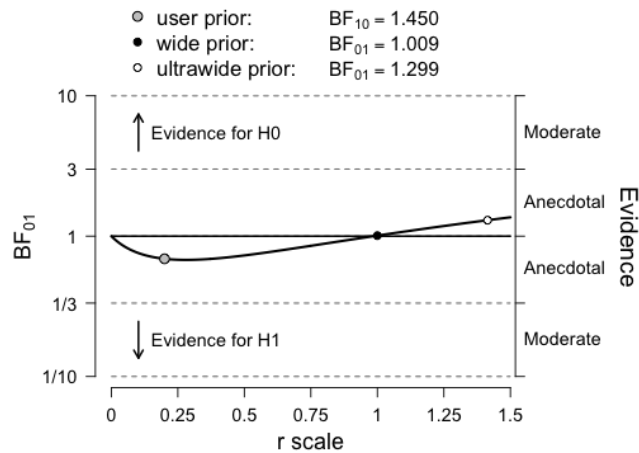


Sample 5, Males: **negative** T×C interaction, $BF_{01} = 1.647$



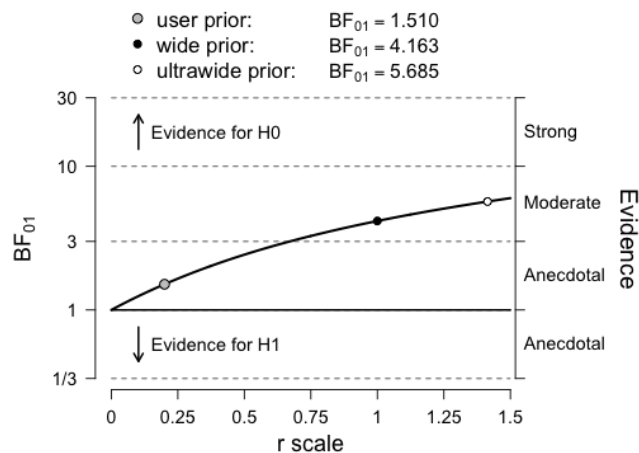
1g. Dominance

Sample 7, Males: **positive** T×C interaction, $BF_{01} = 0.689$



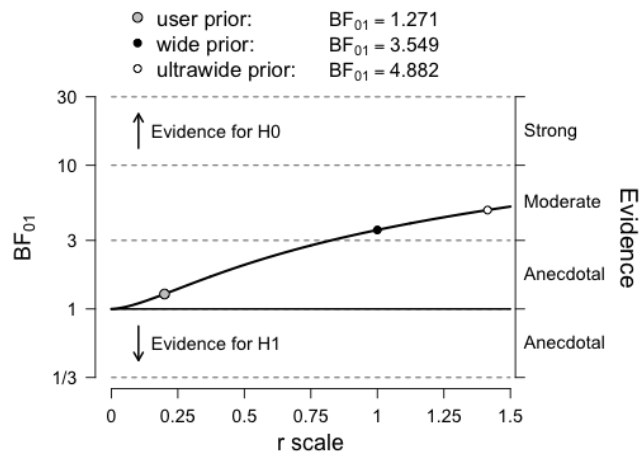
1h. Prestige

Sample 7, Males: **negative** T×C interaction, $BF_{01} = 1.510$

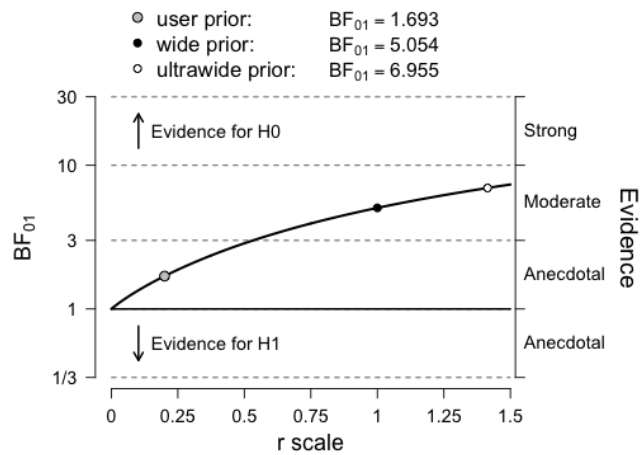


1i. Agreeableness (reversed)

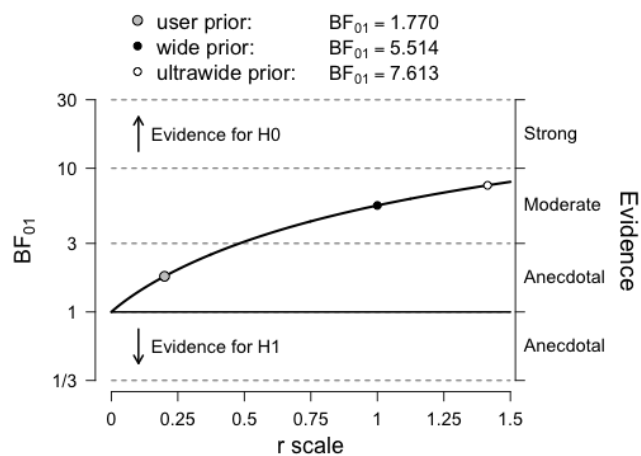
Sample 1, Males: **positive** T×C interaction, $BF_{01} = 1.271$



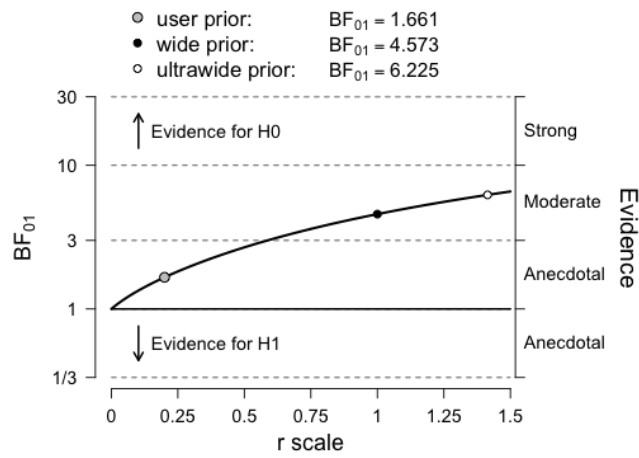
Sample 2, Males: **negative** T×C interaction, $BF_{01} = 1.693$



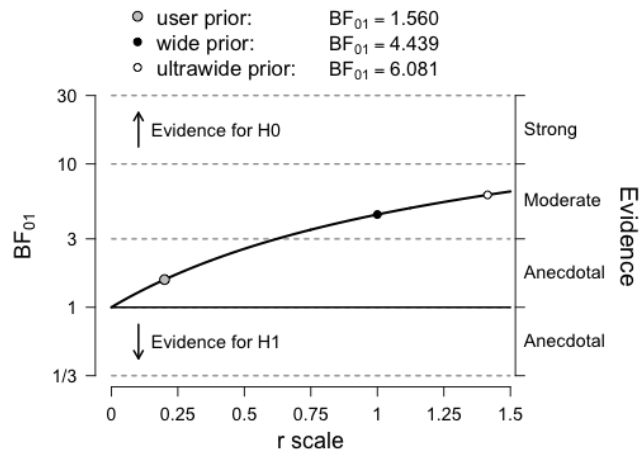
Sample 2, Females: **negative** T×C interaction, $BF_{01} = 1.770$



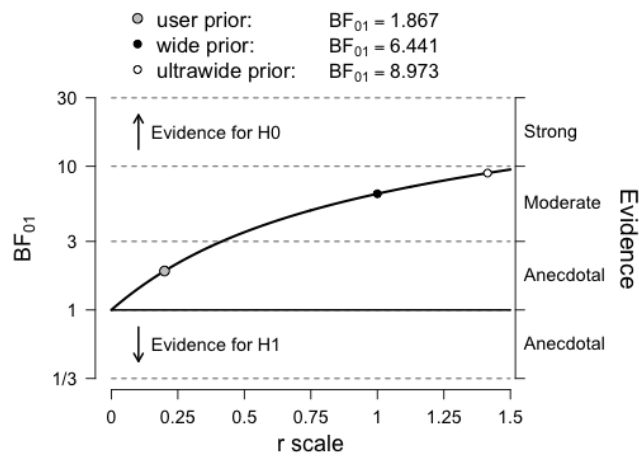
Sample 4, Males: **negative** T×C interaction, $BF_{01} = 1.661$



Sample 4, Females: **positive** T×C interaction, $BF_{01} = 1.560$

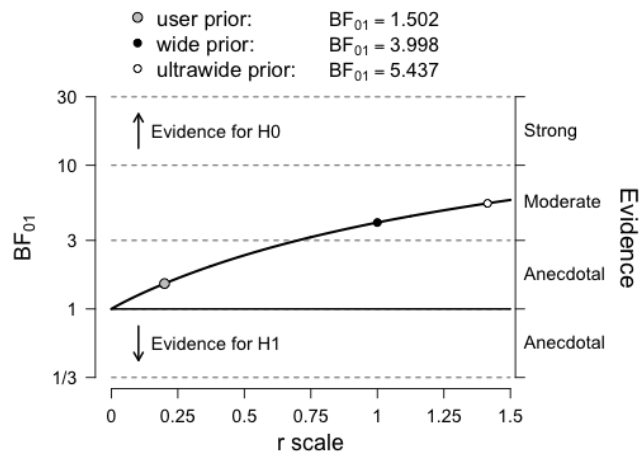


Sample 5, Males: **negative** T×C interaction, $BF_{01} = 1.867$

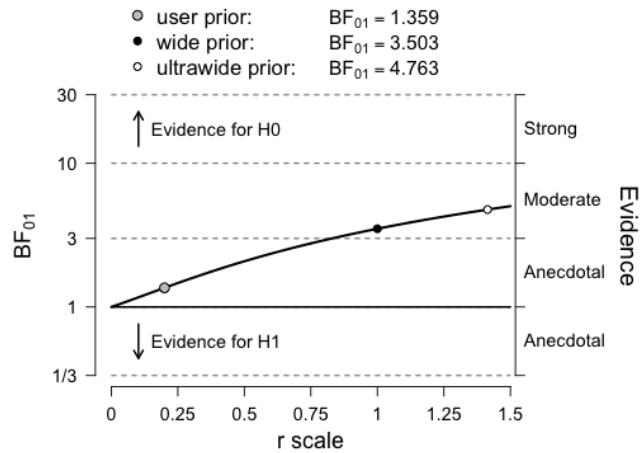


1j. Venturesomeness (EIQ)

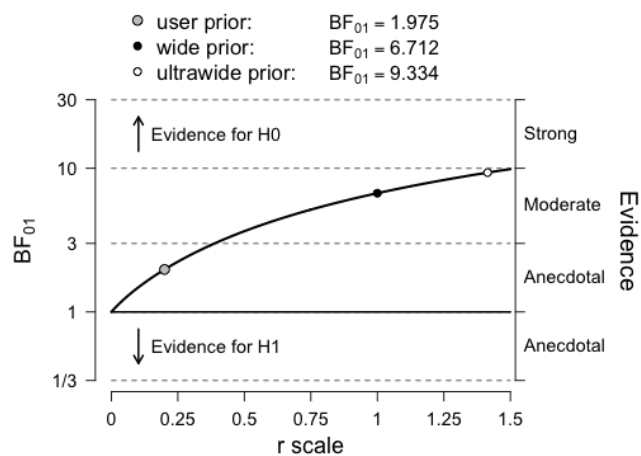
Sample 4, Males: **negative** T×C interaction, $BF_{01} = 1.502$



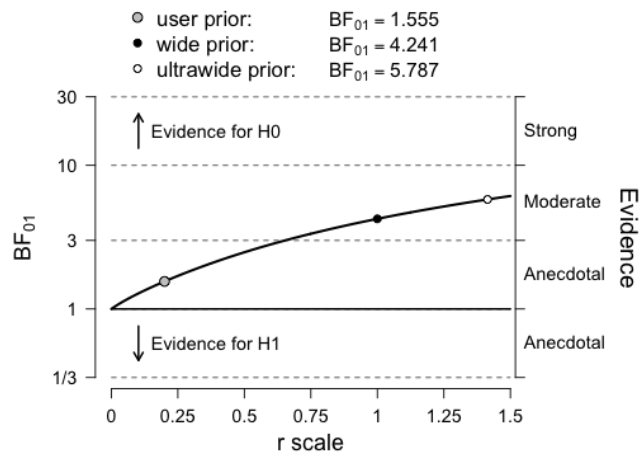
Sample 4, Females: **positive** T×C interaction, $BF_{01} = 1.359$



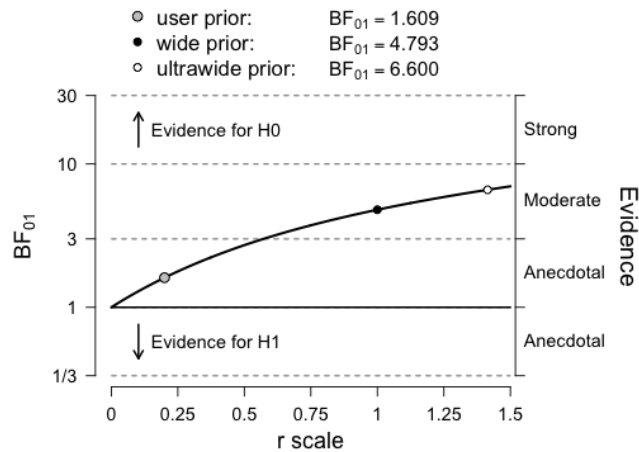
Sample 5, Males: **positive** T×C interaction, $BF_{01} = 1.975$



Sample 6, Males: **negative** T×C interaction, $BF_{01} = 1.555$



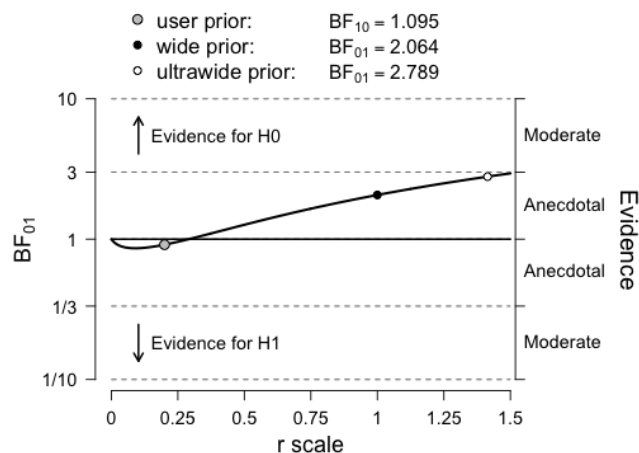
Sample 6, Females: **negative** T×C interaction, $BF_{01} = 1.609$



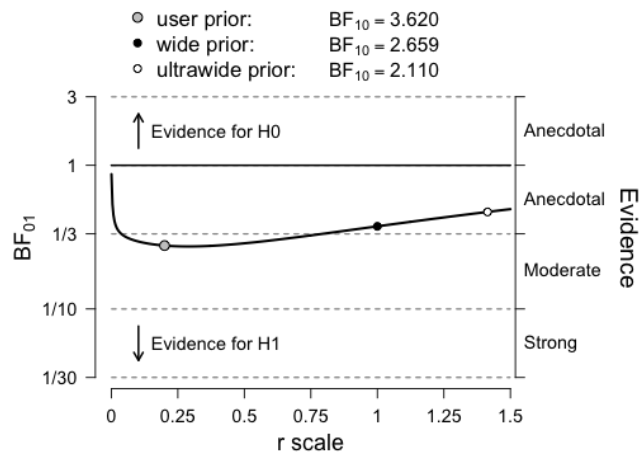
2. Secondary Traits

2a. Extraversion

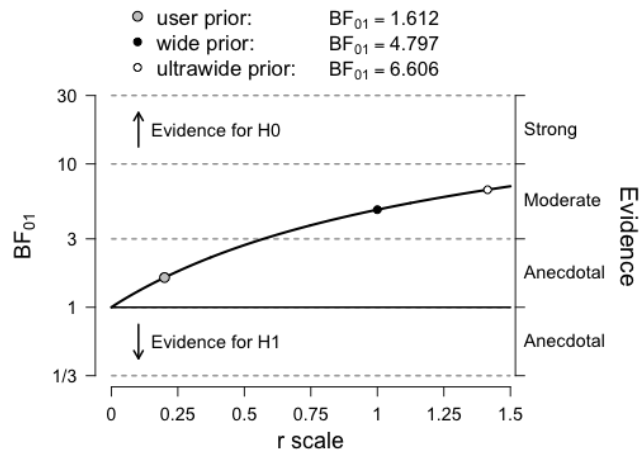
Sample 1, Males: **negative** T×C interaction, $BF_{01} = 0.913$



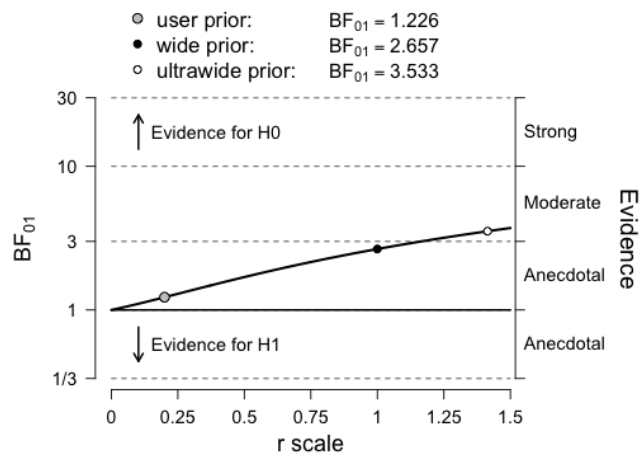
Sample 2, Males: **positive** T×C interaction, $BF_{01} = 0.276$



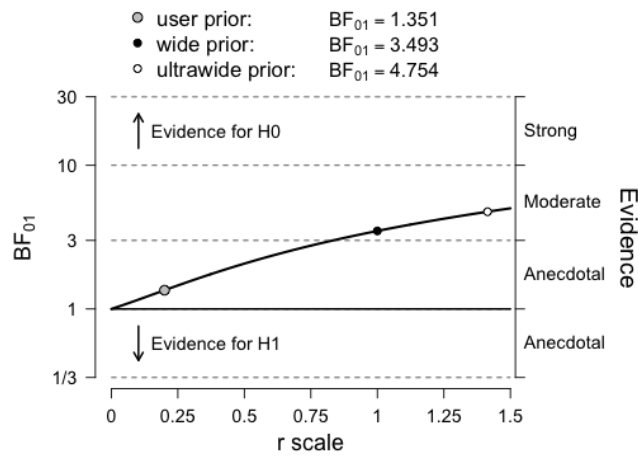
Sample 2, Females: **negative** T×C interaction, $BF_{01} = 1.612$



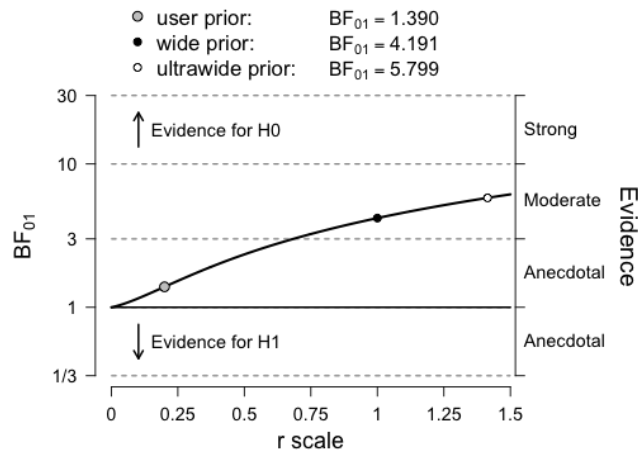
Sample 4, Males: **positive** T×C interaction, $BF_{01} = 1.226$



Sample 4, Females: **positive** T×C interaction, $BF_{01} = 1.351$

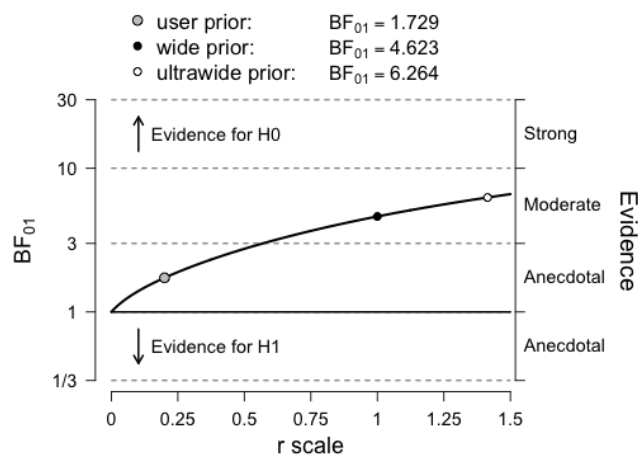


Sample 5, Males: **negative** T×C interaction, $BF_{01} = 1.390$

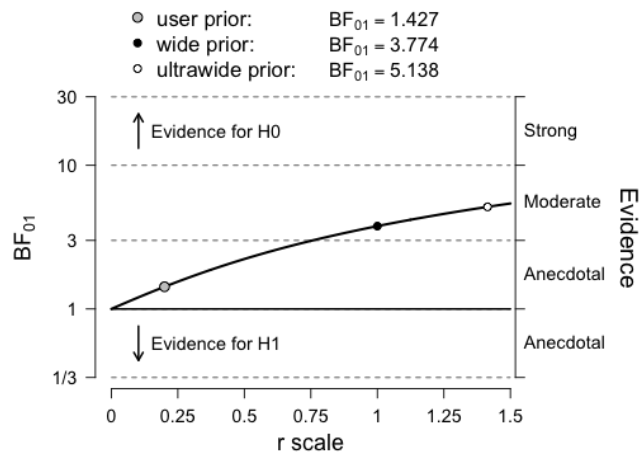


2b. Impulsivity (EIQ)

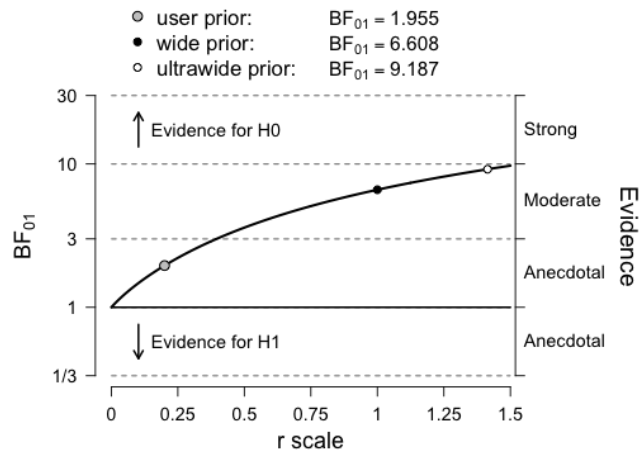
Sample 4, Males: **negative** T×C interaction, $BF_{01} = 1.729$



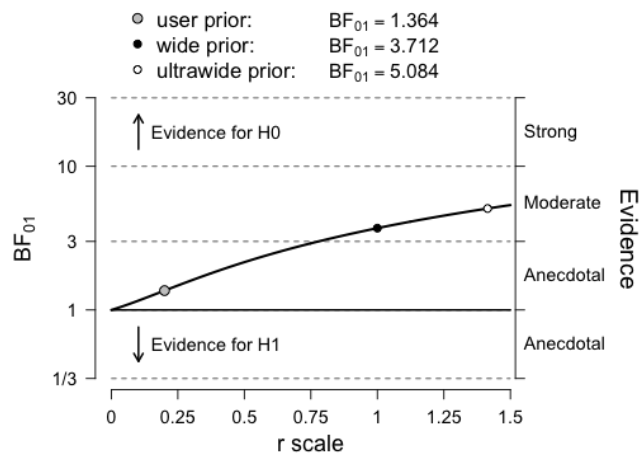
Sample 4, Females: **positive** T×C interaction, $BF_{01} = 1.427$



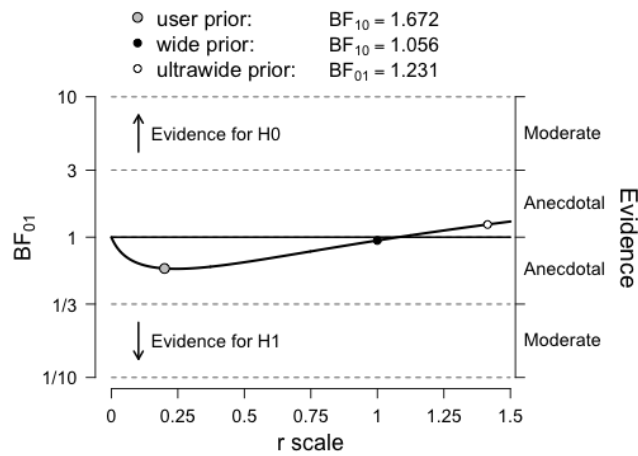
Sample 5, Males: **negative** T×C interaction, $BF_{01} = 1.955$



Sample 6, Males: **negative** T×C interaction, $BF_{01} = 1.364$

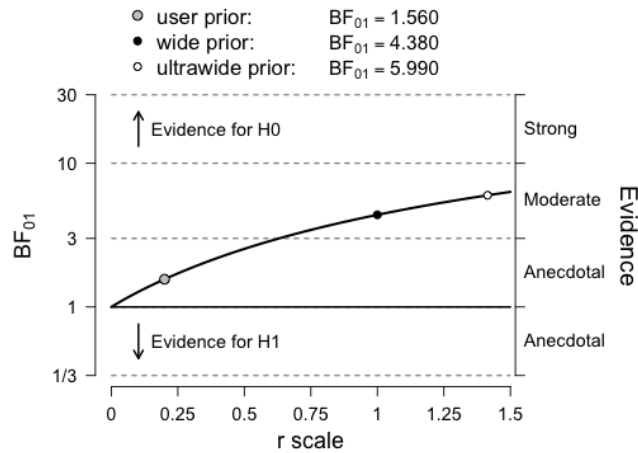


Sample 6, Females: **positive** T×C interaction, $BF_{01} = 0.598$

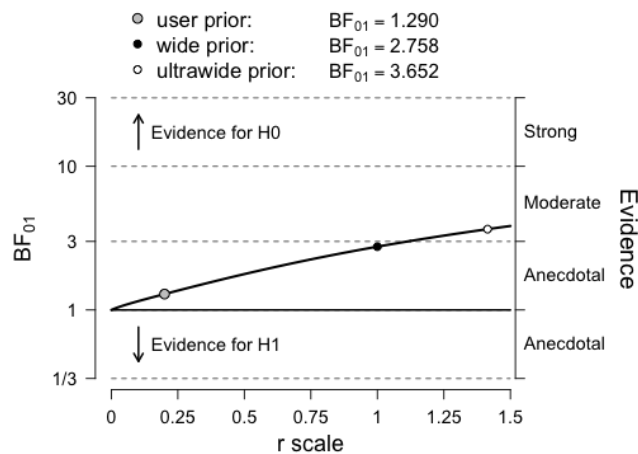


2c. ZTPI present-hedonistic

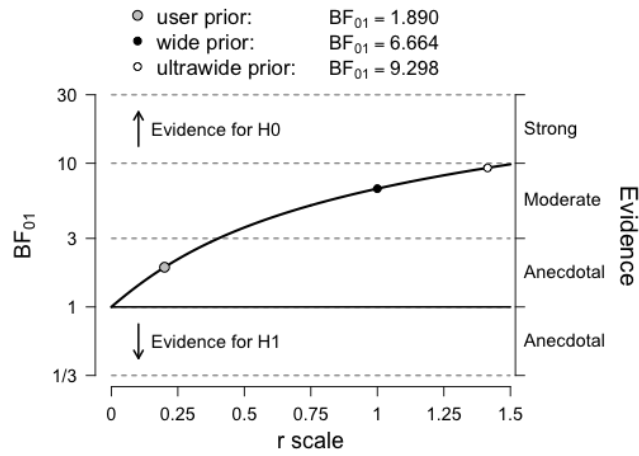
Sample 4, Males: **negative** T×C interaction, $BF_{01} = 1.560$



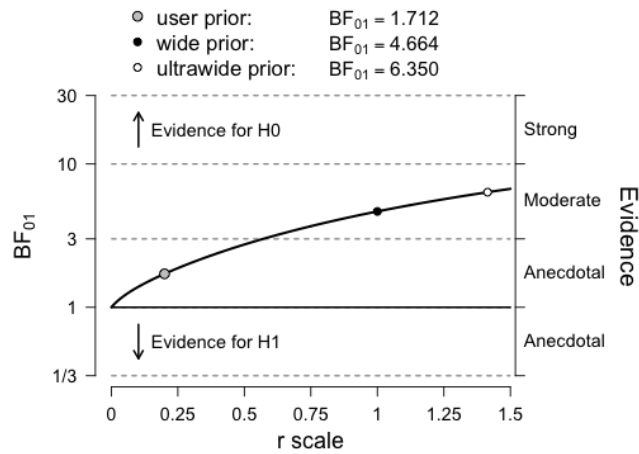
Sample 4, Females: **positive** T×C interaction, $BF_{01} = 1.290$



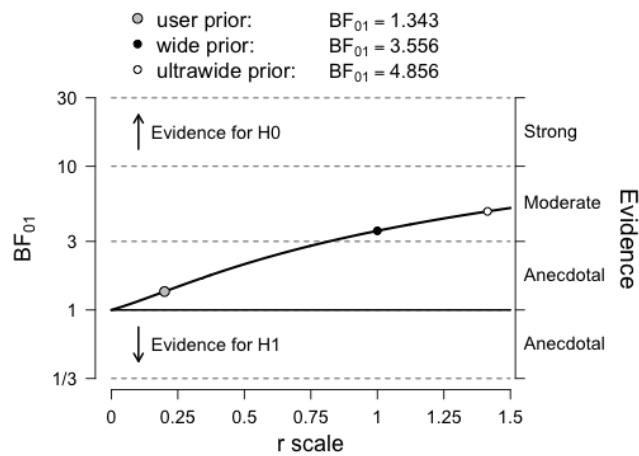
Sample 5, Males: **negative** T×C interaction, $BF_{01} = 1.890$



Sample 6, Males: **positive** T×C interaction, $BF_{01} = 1.712$

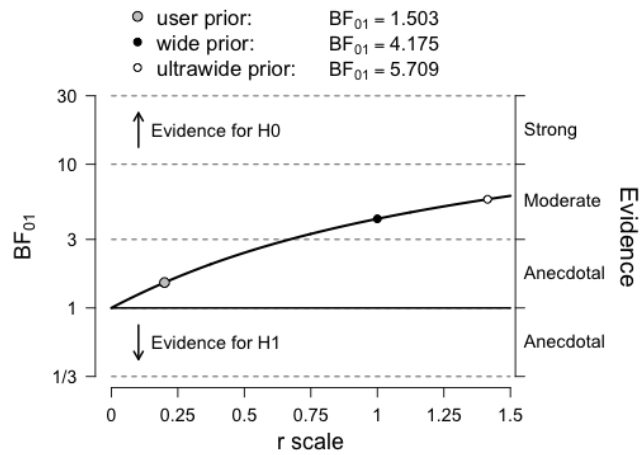


Sample 6, Females: **positive** T×C interaction, $BF_{01} = 1.343$

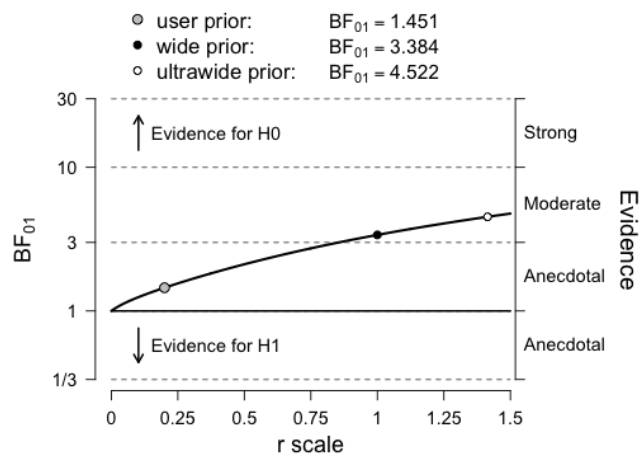


2d. ZTPI present-fatalistic

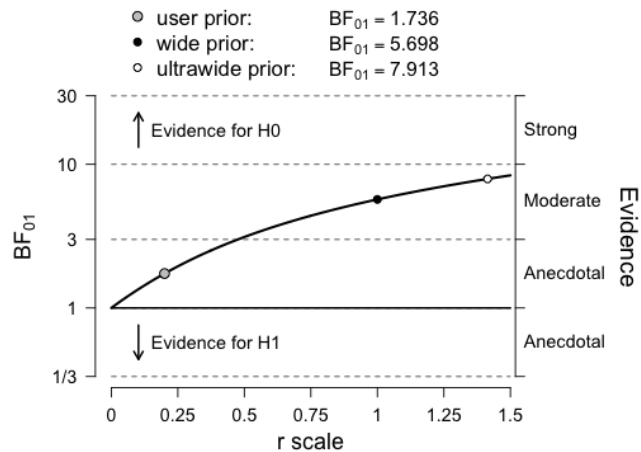
Sample 4, Males: **negative** T×C interaction, $BF_{01} = 1.503$



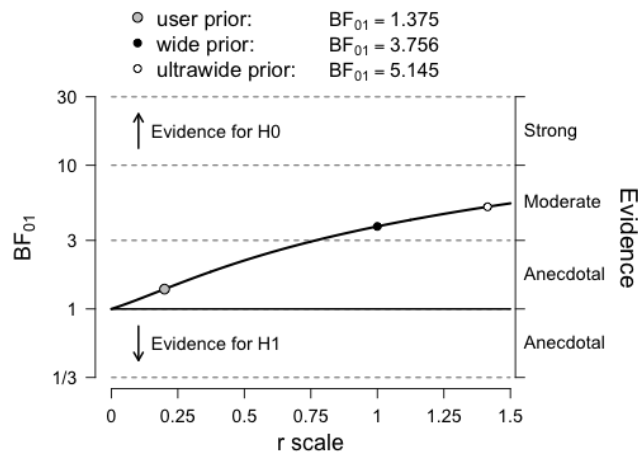
Sample 4, Females: **negative** T×C interaction, $BF_{01} = 1.451$



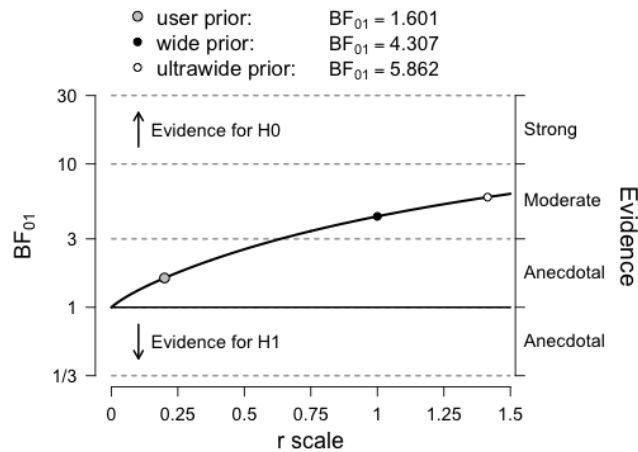
Sample 5, Males: **negative** T×C interaction, $BF_{01} = 1.736$



Sample 6, Males: **positive** T×C interaction, $BF_{01} = 1.375$

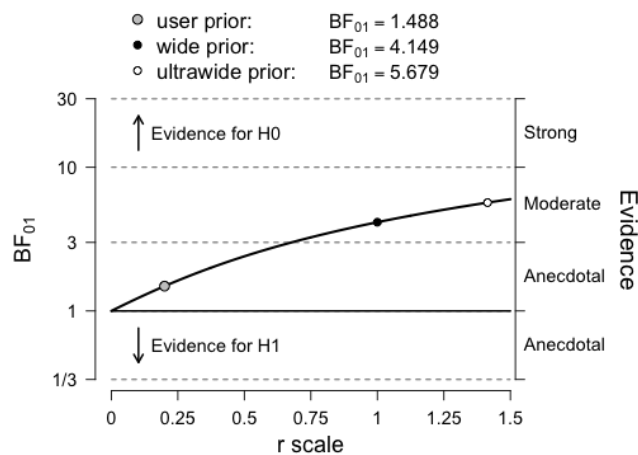


Sample 6, Females: **positive** T×C interaction, $BF_{01} = 1.601$

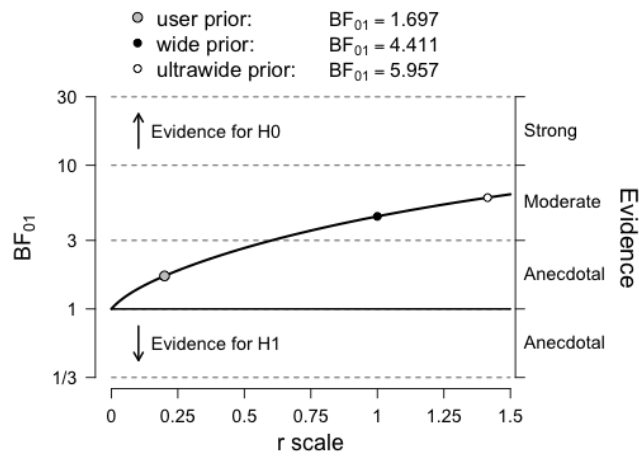


2e. ZTPI future

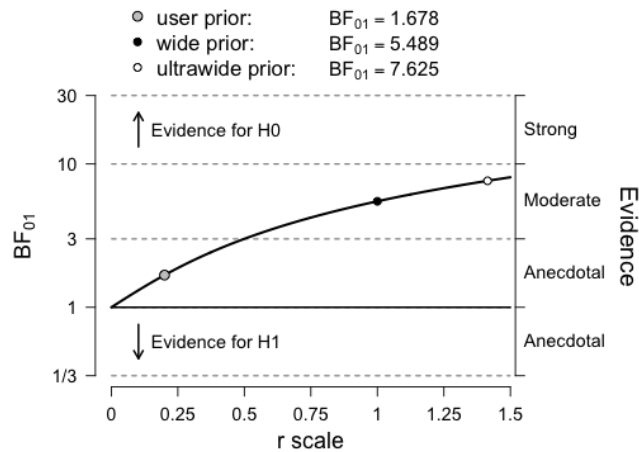
Sample 4, Males: **negative** T×C interaction, $BF_{01} = 1.488$



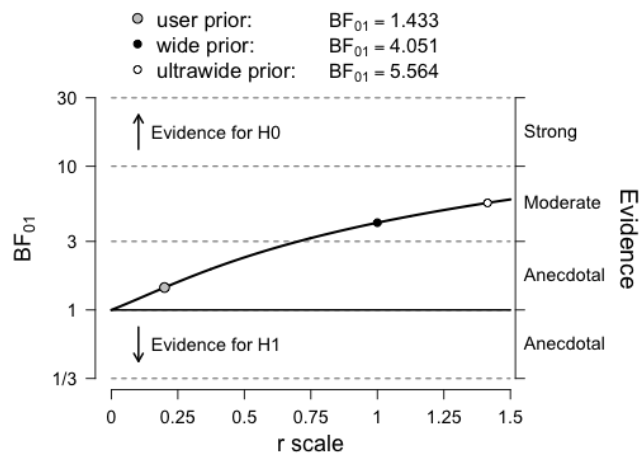
Sample 4, Females: **positive** T×C interaction, $BF_{01} = 1.697$



Sample 5, Males: **negative** T×C interaction, $BF_{01} = 1.678$



Sample 6, Males: **positive** T×C interaction, $BF_{01} = 1.433$



Sample 6, Females: **positive** T×C interaction, $BF_{01} = 1.560$

