Official User-Guide to the *P*-curve

First version: 2013 04 22
This version: 2015 03 02

Uri Simonsohn
Leif Nelson
Joe Simmons

Four steps to a valid *p*-curve:

1. **Create and report a study-selection rule**
2. **Create a *P*-curve Disclosure Table to select results to analyze**
3. **Feed statistical results to *p*-curve app**
4. **Copy-paste app's output onto your paper**

Step 1. Create a study-selection rule

P-curve can be used to assess the evidential value of diverse sets of findings.

If a rule can be specified that creates a meaningful set of studies, then *p*-curve can validly assess the set's joint evidential value.

The rule should be set in advanced, before statistical results are analyzed, and disclosed in the paper.

Examples of rules:

- The yearly top-5 most cited articles in the *Quarterly Research Journal* 1984-1989
- All studies published in 2009 with wine as a manipulation and simulated driving behavior as a dependent variable.
- The most recent 10 articles published by proctologist Giordano Armani.
- Clinicaltrials.gov registered studies examining antidepressants among teenagers.

Step 2. Create a *P*-curve Disclosure Table to select results to analyze

Table 1 summarizes the steps for creating a disclosure table.
Table 2 provides an example.

Table 1.Five Steps to Create a *P*-curve Disclosure Table

Step 1	Identify researchers' stated hypothesis and study design quoting from paper	(Columns 1 &2)
Step 2	Identify the statistical result testing stated hypothesis using Table 3	(Column 3)
Step 3	Report the statistical results of interest quoting from paper	(Column 4)
Step 4	Recompute the precise *p*-value(s) based on reported test statistics	(Column 5)
Step 5	Report robustness results	(Column 6)

Table 2. A Sample *P*-curve Disclosure Table

Original Paper *(Study 1 of each paper)*	**(1) Quoted text from original paper indicating prediction of interest to researchers**	**(2) Study design**	**(3) Key statistical result** *(looking up column 2 in table 3)*	**(4) Quoted text from original paper with statistical results**	**(5) Results**	**(6) Robustness results**
Van Boven et al. (2010)	We predicted that people would perceive their embarrassing moment as less psychologically distant when described emotionally.	**two-cell (description: emotional vs. not)**	**Difference of means**	As predicted, participants perceived their previous embarrassing moment to be less psychologically distant after describing it emotionally (M = 4.90, SD = 2.30) than after describing it neutrally (M = 6.66, SD = 1.83), **$t(38) = 2.67$, $p < .025$** (see Table 1).	t(38)=2.67, p=0.0111	
Topolinksi & Strack (2010)	We predicted that the classical effect by Jacoby, Kelley, et al., namely the **misattribution** of increased fluency to fame, **would vanish** under the oral motor task but would still be detected under a manual motor task.	**2 (exposure: old vs. new) x 2 (motor task: oral vs manual)** *(attenuated interaction)*	**Two-way interaction**	Over the fame ratings in the test phase, a 2 (exposure: old items, new items) × 2 (concurrent motor task: manual, oral) analysis of variance (ANOVA) was run with motor task as a between-subjects factor. A main effect of exposure, F(1, 48) = 5.54, p < .023, ηp2 = .10, surfaced, as well as an interaction between exposure and motor task, **$F(1, 48) = 4.12$, $p < .05$, ηp2 = .08.** The conditional means are displayed in Table 1.	F(1,48)=4.12, p=0.0479	
Clarkson et al. (2010)	Specifically, participants in the low depletion condition were expected […] to **persist longer** on our problem-solving task when given the replenished (vs. depleted) feedback. **Conversely,** participants in the high depletion condition were expected […] to persist longer on our problem-solving task when given the depleted (vs. replenished) feedback.	**2 (depletion: high vs low) x 2 (feedback: depleted vs. replenished)** *(reversing interaction)*	**Two simple effects**	In the low depletion condition, participants persisted significantly longer when given the replenished, as opposed to depleted, feedback, **$t(30) = -2.52$, $p < .02$.** In the high depletion condition, participants persisted significantly longer when given the depleted, as opposed to replenished, feedback, **$t(30) = 2.50$, $p < .02$.**	t(30)=2.52, p=0.0173 t(30)=2.5, p=0.0181	
Wohl & Branscombe (2005)	We expected that Jews would be more willing to **forgive** Germans for the past when they categorized at the human identity level and that the **guilt assigned** to contemporary Germans would be lower in the human identity condition compared with the social identity condition.	**two-cell (identity: human vs. social)**	**Difference of means** *(for two d.v.s)*	Participants assigned significantly less collective guilt to Germans when the more inclusive human-level categorization was salient (M = 5.47, SD = 2.06) than they did when categorization was at the social identity level (M = 6.75, SD = 0.74), **F(1, 45) = 7.62, p <.01**, d = 0.83. Participants were more willing to **forgive** Germans when the human level of identity was salient (M = 5.84, SD = 1.25) than they were when categorization was at the social identity level (M = 4.52, SD = 0.92), **F(1, 45) = 16.55, p <.01, d = 1.20.**	F(1,45)=7.62, p=0.0083	F(1,45)=16.55, p=0.0002

Column (1) in Table 2 includes the text that identifies the appropriate statistical result to select. For example:

- Topolinski & Strack (2010) write that the effect is expected to "vanish," so they predict an attenuating interaction. Table 3 below indicates that for attenuating interactions one selects the statistical results associated with the interaction.

- Clarkson's et al. (2010) expect the effect to reverse in sign across conditions, so they predict a reversing interaction. Table 3 below indicates that for reversing interactions one selects the statistical result associated with both simple effects.

Step 2 (cont.)

Table 3 in paper. Which statistical result to select for *p*-curve?

DESIGN	EXAMPLE	WHICH RESULT TO INCLUDE IN MAIN P-CURVE	WHICH RESULT TO INCLUDE IN ROBUSTNESS TEST
3-Cell	*Examining how math training affects math performance*		
High **Medium** **Low**	60 minutes of math training 30 minutes of math training 5 minutes of math training	Linear trend	High vs. low comparison
Treatment **Control 1** **Control 2**	60 minutes of math training 60 minutes of unrelated training No training	Treatment vs. Control 1	Treatment vs. control 2
Treatment 1 **Treatment 2** **Control**	60 minutes of math training, start with easy questions 60 minutes of math training, start with hard questions No training	Treatment 1 vs. Control	Treatment 2 vs. Control
2X2 DESIGN	*Examining how season interacts with being indoors vs. outdoors to affect sweating*		
Attenuated Interacton	Always sweat more in summer, but less so when indoors.	2x2 Interaction	
Reversing Interacton	Sweat more in summer than winter when outdoors, but more in winter than in summer when indoors	Both simple effects	
3x2 DESIGN	*Examining how season interacts with math training to affect math performance*		
Attenuated Trends	More math training (60 vs. 30 vs. 5 minutes) leads to better performance always, but more so in winter than in summer	Difference in linear trends	2x2 interaction for high vs. low
Reversing Trends	More math training (60 vs. 30 vs. 5 minutes) leads to better performance in winter, but worse performance in summer	Both linear trends	Both high vs. low comparisons
2x2x2 DESIGN	*Examining how gender and season interact with being indoors vs. outdoors to affect sweating*		
***Attenuation* of attenuated interaction**	Both men and women sweat more in summer than winter, but less so when indoors. This attenuation is stronger for men than for women.	Three-way interaction	
***Reversal* of attenuated interaction**	Men sweat more in summer than winter, but less so when indoors. Women also sweat more in summer than winter, but more so when indoors.	Both two-way interactions	
***Reversal* of reversing interaction**	Men sweat more in summer than winter when outdoors, but more in winter than in summer when indoors. Women sweat more in winter than summer when outdoors, but more in summer than winter when indoors.	All four simple effects	

Keep in mind:

Important!

1. **In a 2x2 design,**
   - If attenuation is predicted, select only the interaction
   - If a reversal is predicted, select only both simple effects

2. ***Discrete tests.***
   P-curve is only approximately valid for discrete tests (e.g., comparing proportions). *P*-curves of discrete tests are, for now, merely suggestive. See Supplement #4.

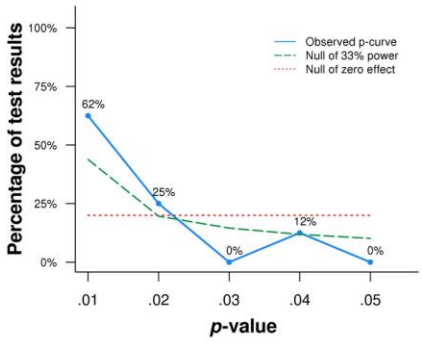
Step 3. Feed key results to *p*-curve app (version 3.0)

The web-based app looks like this:

You can copy paste your tests in the format used in the examples there.
If you have results p>.05, the app will automatically exclude them and report how many were excluded.

Step 4. Copy-paste app's output onto your paper *(or email/tweet/blogpost)*

After clicking on you will see a screen like this one:



Statistical Inference	Results	
	Binomial Test (Share of significant results p<.025)	Continuous Test (Aggregate pp-values via Stouffer Method)
1) Studies contain evidential value. (Right skew)	p =.0352	Z = -3.94, p<.0001
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p =.9224	Z = 1.83, p=.9663
3) Studies exhibit evidence of intense p-hacking. (Left skew)	p =.9961	Z = 3.94, p>.9999
Estimate of Statistical Power		
Average power of tests included in p-curve (correcting for publication bias)	71%	

The observed p-curve includes 8 significant results (p<.05), of which 87.5% are p<.025.
There were no non-significant results entered.

If you right-click on the figure itself you can save it as an image file, but you will not save the text below it.

To grab the entirety of the output, as done above, you can do a printscreen.

If you haven't done that before, check these instructions out for Windows 7 or XP or Mac. If you use a Unix machine you probably have not read this far.