Assessing power and effect sizes in the recent Cognitive Translation Studies literature (2009-2018)

Dr. Christian Olalla-Soler

Universitat Autònoma de Barcelona

TREC

Index

1. Sample size, effect size significance and power
2. Research questions
3. Power assessment
4. Results
    1. Global sample size, effect size and power
    2. Differences in sample size, effect size and power between significant and non-significant tests
    3. Sample size, effect size and power for *d*, *f* and ρ
    4. Lower cut-off values for low, medium and high effect sizes for *d*, *f* and ρ
    5. Evolution of sample size, effect size and power
5. Conclusions

Sample size, effect size significance and power

"The power of a statistical test is the probability that it will yield statistically significant results" (Cohen, 1988: 1)

Significance (α): the lower the value, the lower the power of a test. Unilateral tests result in higher power.

Sample size (n): the larger the sample size, other things being equal, the smaller the error and the greater the reliability or precision of the results.

Effect size (ES): "the degree to which the phenomenon is present in the population" (Cohen, 1988: 9). The larger the effect size, the greater the power of the test. The larger the effect size, the smaller the sample size necessary to detect it.

Research questions

1. What is the median sample size, effect size and power in Cognitive Translation Studies (CTS)?
2. What differences are there regarding sample size, effect size and power between significant and non-significant tests?
3. What sample size is needed in CTS to reach a median low, medium and large effect size threshold at power = 0.80 and α =0.05?
4. How have sample size, effect size and power evolved in the last 10 years (2009-2018)?

Power assessment

1. Power as a function of ES, α and n. Used for planning research and for determining the post-hoc power of a test.
2. n as a function of power, ES and α. Used for determining the sample size. A researcher wants to know what n should s/he have to have X power to detect Y effect size at Z α level.
3. ES as a function of α, n and power. Used for determining the ES that can be detected for given α, n and power.
4. α as a function of ES, n and power. What significance level must I use to detect a given ES with specified power for a fixed given n?

Power assessment

- Manual extraction of sample sizes, effect sizes and α levels from CTS publications from 2009-2018.
- All effect sizes included in the assessment were converted to *d*.
- Computation of power using G*Power v.3 (Faul et al., 2007).
- Computation of descriptive statistics (mean, median, SD, interquartile range).
- Computation of 95% confidence intervals (bootstrapping with 1000 samples; bias-corrected and accelerated).

Power assessment

Initial corpus (BITRA; 2009-2018; CTS)	**632 documents**
Documents that did not include tests	472 (74.7%)
Documents in which it was not posible to compute effect sizes and power	56 (8.9%)
Inaccessible documents	19 (3%)
Documents whose unit of analysis was not participants (e.g. corpus research)	2 (0.3%)
Documents discarded because the tests employed were not representative enough (> 25 tests of the same type for which effect size and power were computable)	18 (2.8%)
Final corpus	**65 documents (517 tests)**

Power assessment

Effect sizes

- *d* (311; 60.2%)
  - Dependent *t*: n = 153 (49.2%)
  - Independent t: n = 84 (27.0%)
  - *U*: n = 74 (23.8%)
- *f* (103; 19.9%)
  - Fixed effects one-way ANOVA: n = 92 (89.3%)
  - ANCOVA: n = 11 (10.7%)
- ρ (103; 19.9%)
  - correlations: n = 103 (19.9%)

Power assessment

- α = 0.05: 477 (92.3%)
- α = 0.025: 40 (7.7%)

Limitations of power assessment

- Power analysis is useless when applied to non-significant results: post-hoc power is a deterministic function of the *p* value. Whenever the obtained *p* value is greater than the α level (e.g., .05), post-hoc power is always less than 50% (Wang, 2010; Gelman, 2019).
- Generally, confidence intervals provide more information than power (Dziak, Dierker and Abar, 2018).
- Calculating power using observed effect sizes: the observed effect is not the true effect; it's just an estimate thereof (Reito and York, 2018).

- OK: power assessment to describe power levels of a discipline.
- OK: estimating real effect sizes through confidence intervals.

Results

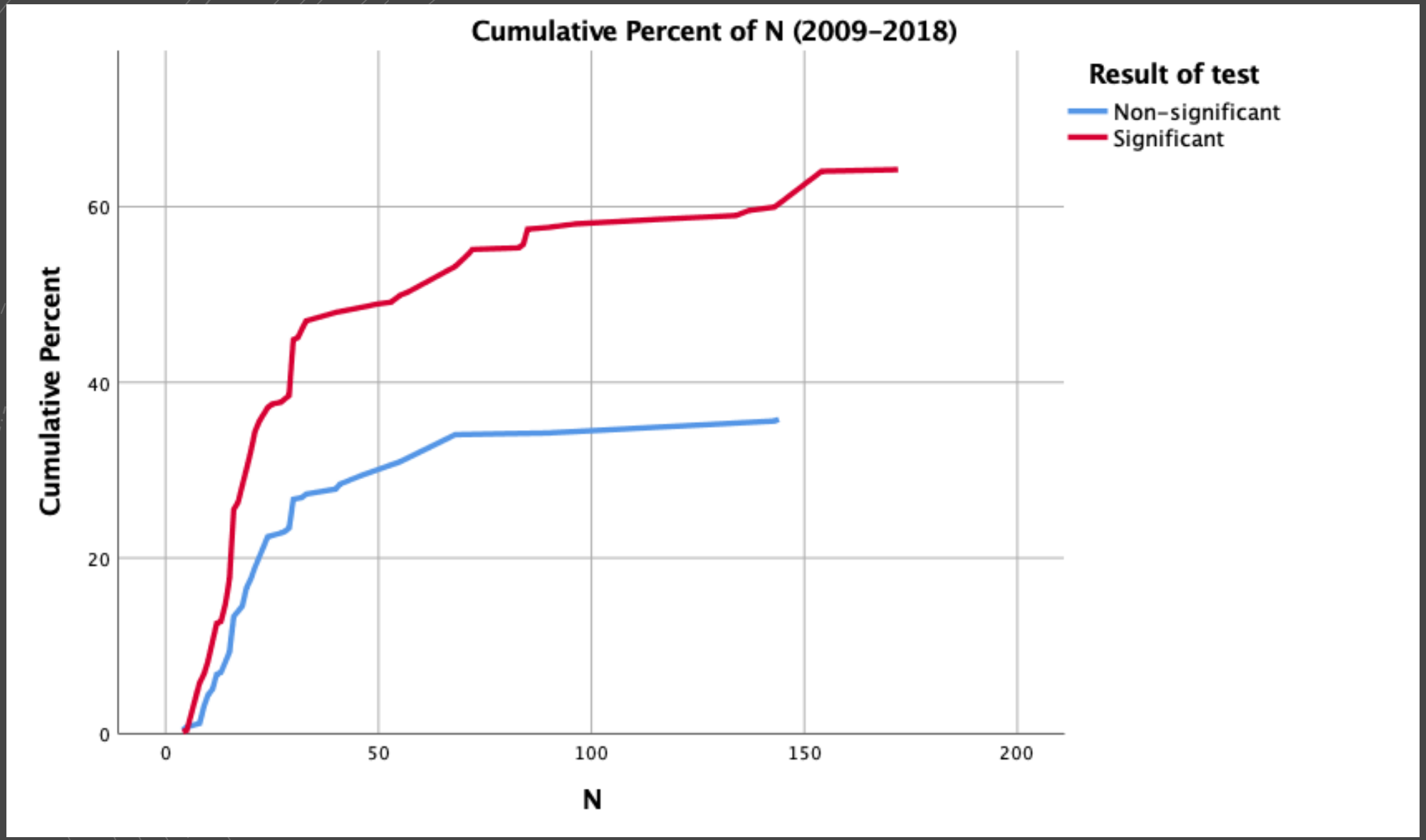
Global sample size, effect size and power

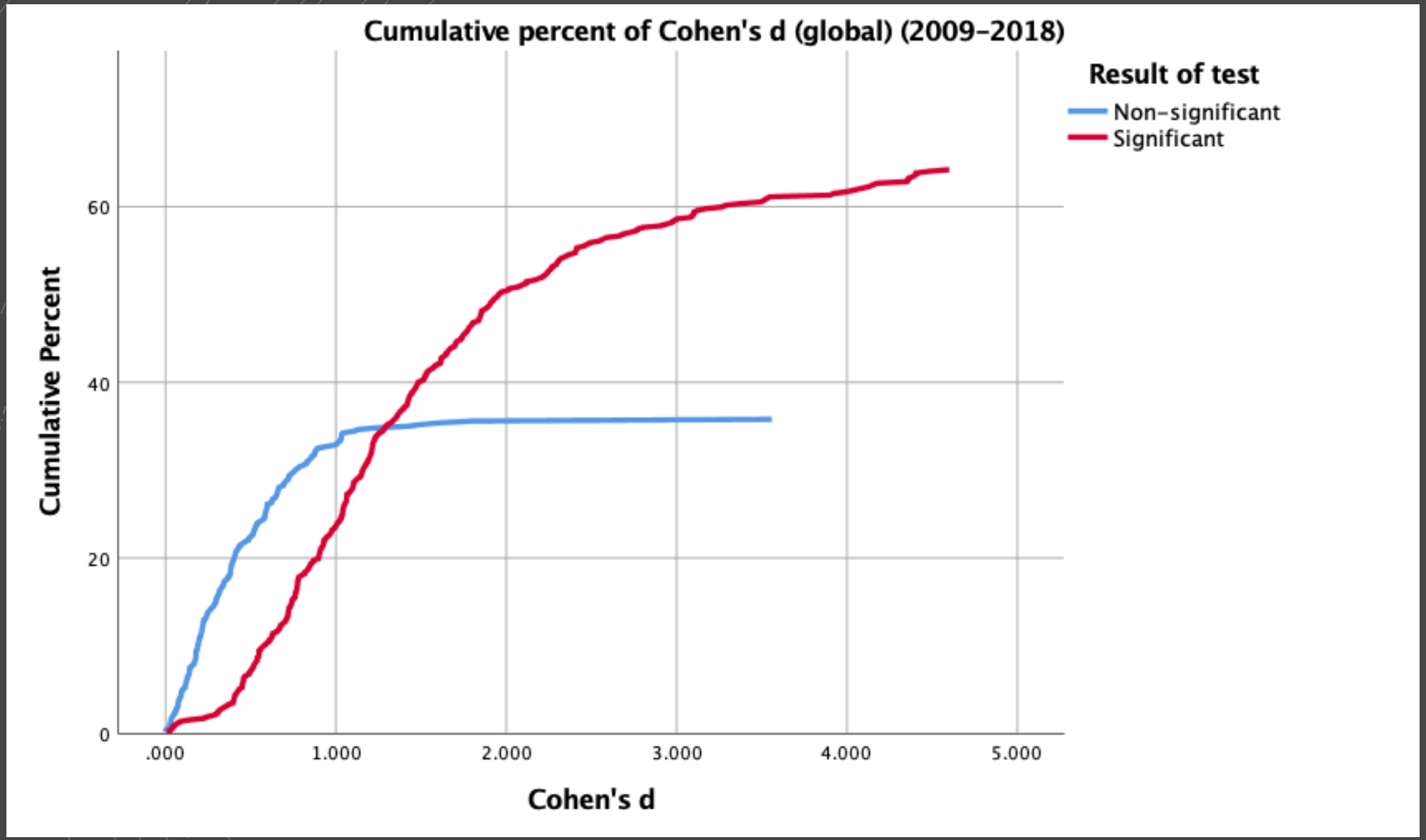
Mean, standard deviation and interquartile range for sample size, global Cohen's *d* and power for the whole period of analysis (2009-2018)

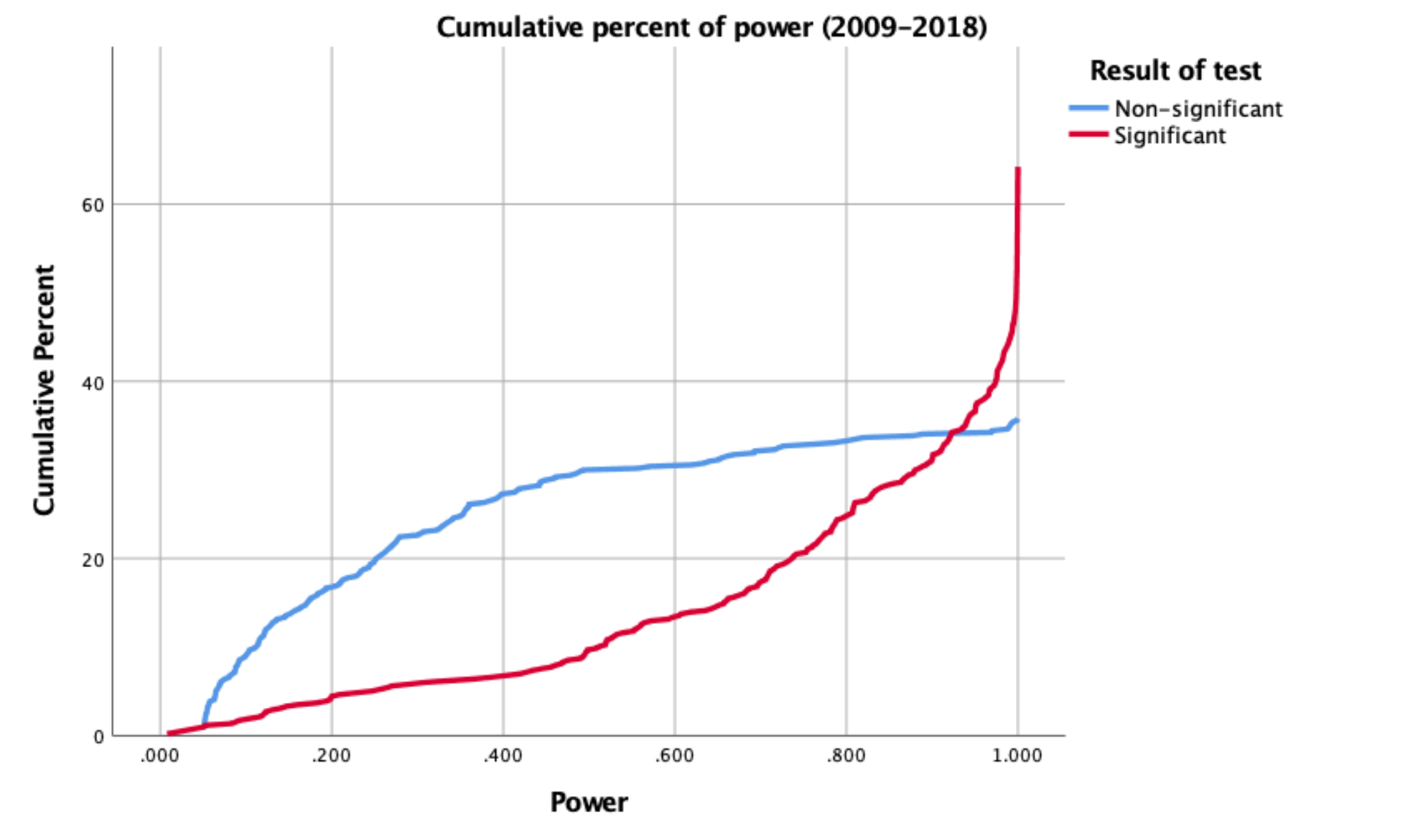
2009-2018 (n = 511)	N	Cohen's *d* (global)	Power
Median	21 95% CI [20-21]	0.840 95% CI [0.762-0.919]	0.706 95% CI [0.651-0.755]
Mean	36.31 95% CI [33.37-39.67]	1.099 95% CI [1.025-1.181]	0.611 95% CI [0.580-0.643]
SD	37.49	0.953	0.351
Interquartile range	25 95% CI [17-30]	1.074 95% CI [0.951-1.212]	0.722 95% CI [0.683-0.764]

Cohen's *d* thresholds (Cohen, 1988): small = 0.2; medium = 0.5; large = 0.8; recommended power (Cohen, 1988) = 0.80

Differences in sample size, effect size and power between significant and non-significant tests


Cumulative Percent of N (2009–2018)
Result of test
Non-significant
Significant
Cumulative Percent
0
20
40
60
0
50
100
150
200
N


Cumulative percent of Cohen's d (global) (2009–2018)
Result of test
Non-significant
Significant
Cumulative Percent
0
20
40
60
.000
1.000
2.000
3.000
4.000
5.000
Cohen's d


Cumulative percent of power (2009–2018)
Result of test
Non-significant
Significant
Cumulative Percent
0
20
40
60
.000
.200
.400
.600
.800
1.000
Power

Sample size, effect size and power for *d*, *f* and ρ

Mean, standard deviation and interquartile range for sample size, Cohen's *d* and power for the whole period of analysis (2009-2018)

2009-2018 (n = 311)	n	Cohen's *d*	Power
Median	18 95% CI [18, 18.50]	0.654 95% CI [0.568, 0.742]	0.461 95% CI [0.373, 0.549]
Mean	26.99 95% CI [24.31, 30.05]	0.846 95% CI [0.754, 0.931]	0.521 95% CI [0.483, 0.557]
SD	27.879	0.793	0.360
Interquartile range	15 95% CI [12, 18]	0.859 95% CI [0.723, 1.036]	0.761 95% CI [0.690, 0.816]

Cohen's *d* thresholds (Cohen, 1988): small = 0.2; medium = 0.5; large = 0.8; recommended power (Cohen, 1988) = 0.80

Mean, standard deviation and interquartile range for sample size, *f* and power for the whole period of analysis (2009-2018)

2009-2018 (n = 103)	n	*f*	Power
Median	68 95% CI [68, 71]	0.420 95% CI [0.369, 0.523]	0.803 95% CI [0.719, 0.888]
Mean	73.71 95% CI [62.83, 82.82]	0.537 95% CI [0.458, 0.628]	0.673 95% CI [0.606, 0.749]
SD	45.56	0.793	0.338
Interquartile range	55 95% CI [52, 55]	0.62 95% CI [0.270, 0.553]	0.630 95% CI [0.443, 0.769]

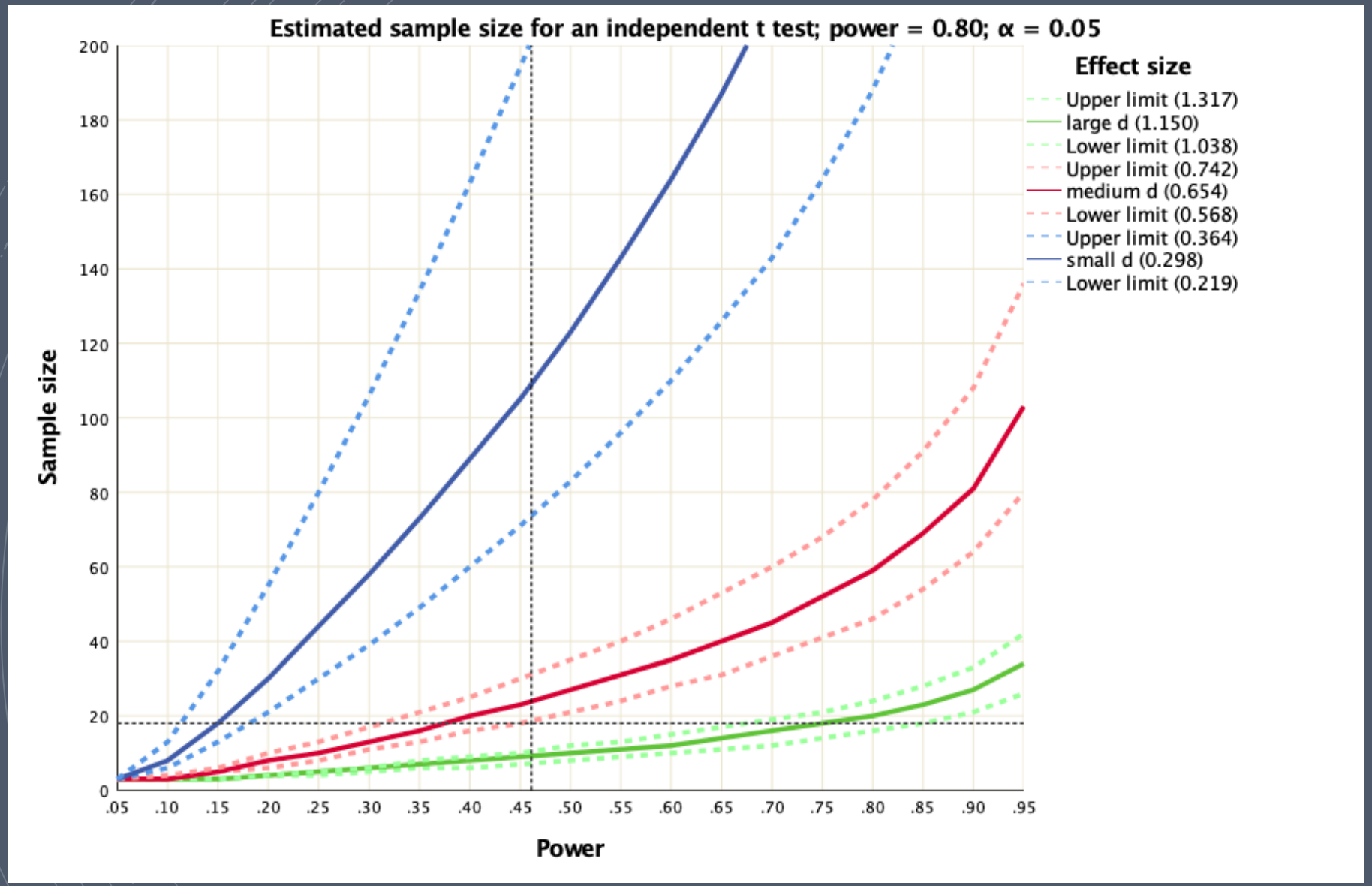
Cohen's *f* thresholds (Cohen, 1988): small = 0.1; medium = 0.25; large = 0.4; recommended power (Cohen, 1988) = 0.80

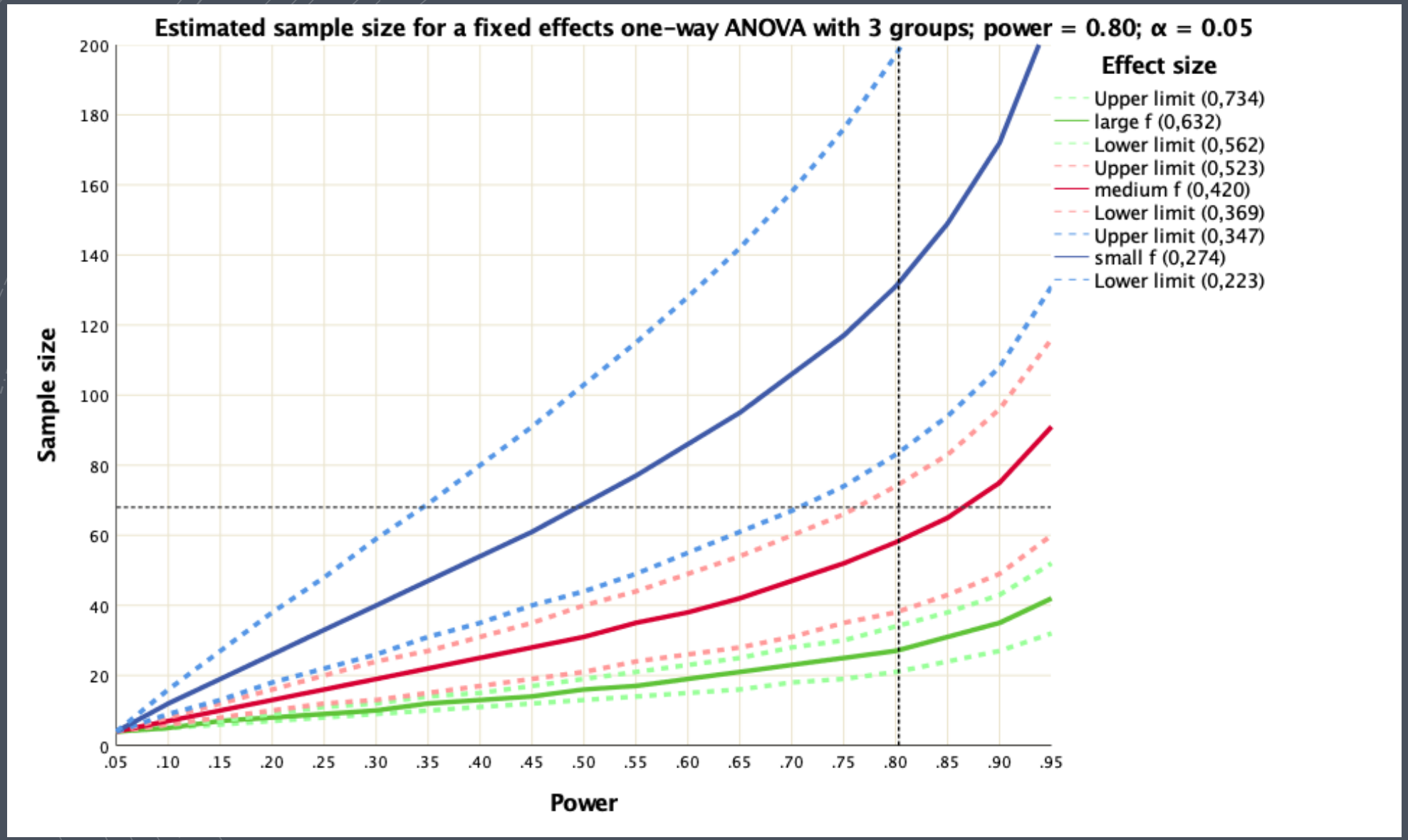
Mean, standard deviation and interquartile range for sample size, ρ and power for the whole period of analysis (2009-2018)

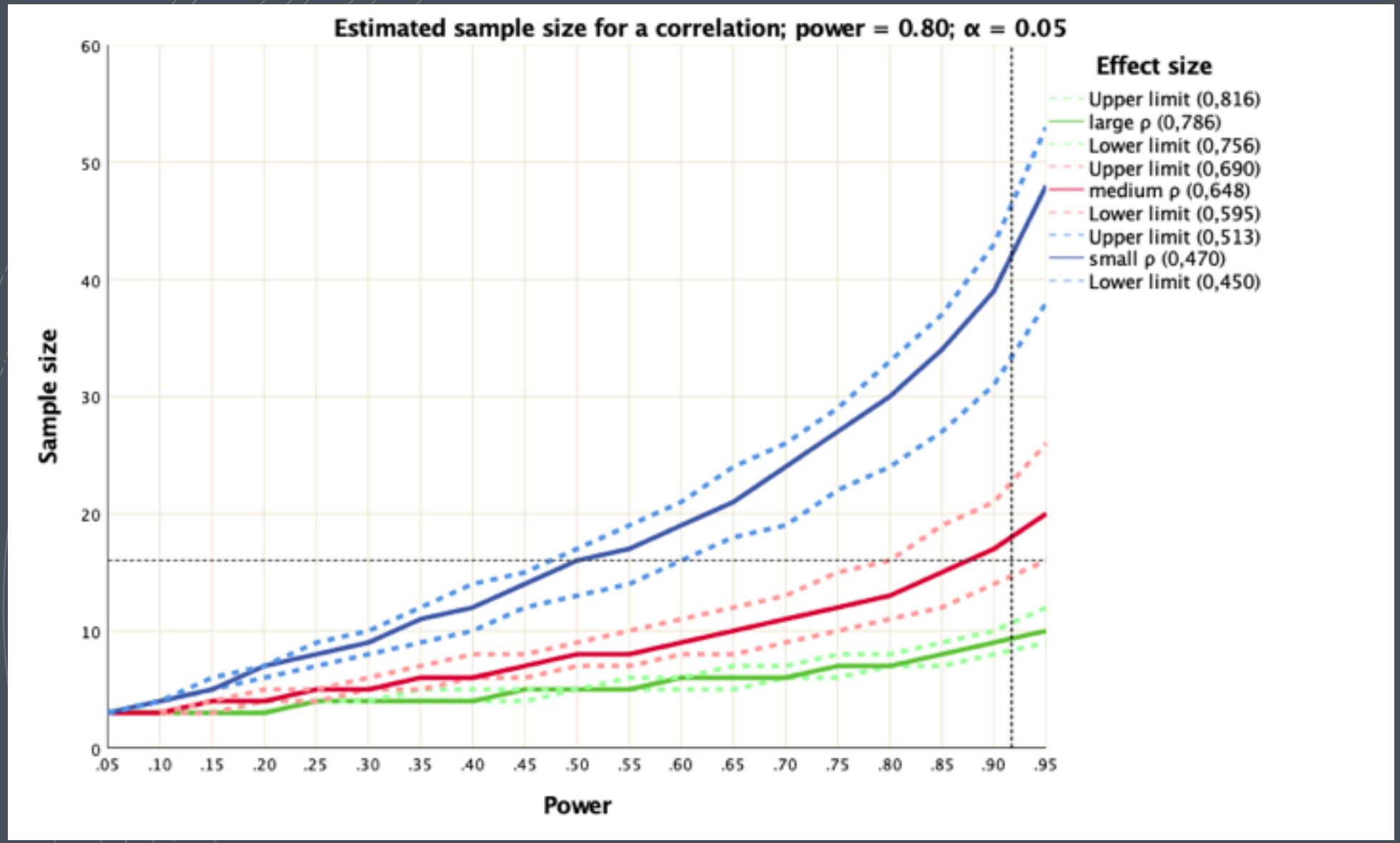
2009-2018 (n = 103)	n	ρ	Power
Median	16 95% CI [16, 16]	0.648 95% CI [0.595, 0.690]	0.917 95% CI [0.857, 0.955]
Mean	27.49 95% CI [22.05, 33.27]	0.618 95% CI [0.577, 0.656]	0.832 95% CI [0.791, 0.875]
SD	31.10	0.209	0.201
Interquartile range	18 95% CI [9, 19]	0.321 95% CI [0.267, 0.361]	0.287 95% CI [0.223, 0.311]

Cohen's *d* thresholds (Cohen, 1988): small = 0.1; medium = 0.3; large = 0.5; recommended power (Cohen, 1988) = 0.80

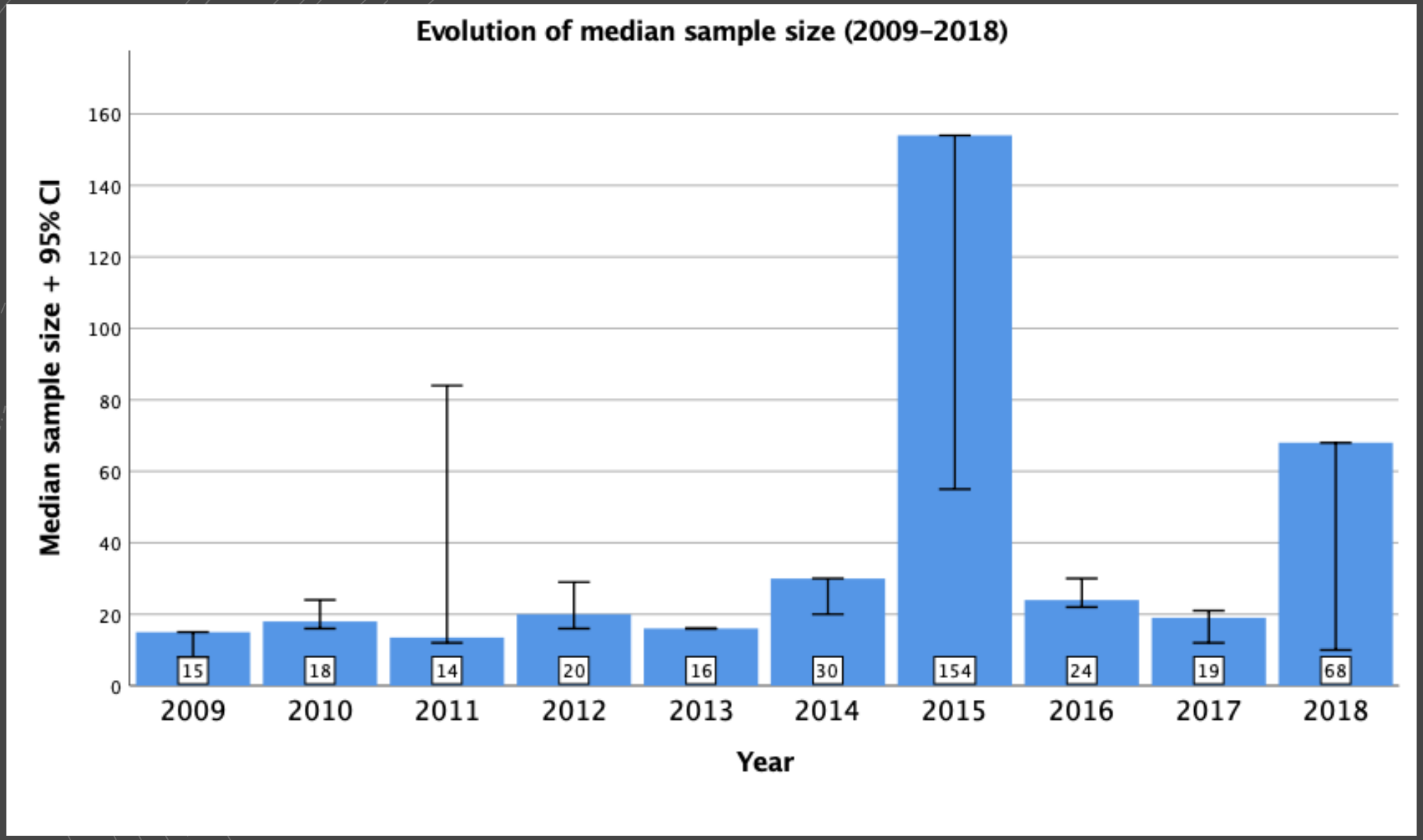
Lower cut-off values for low, medium and high effect sizes for *d*, *f* and ρ

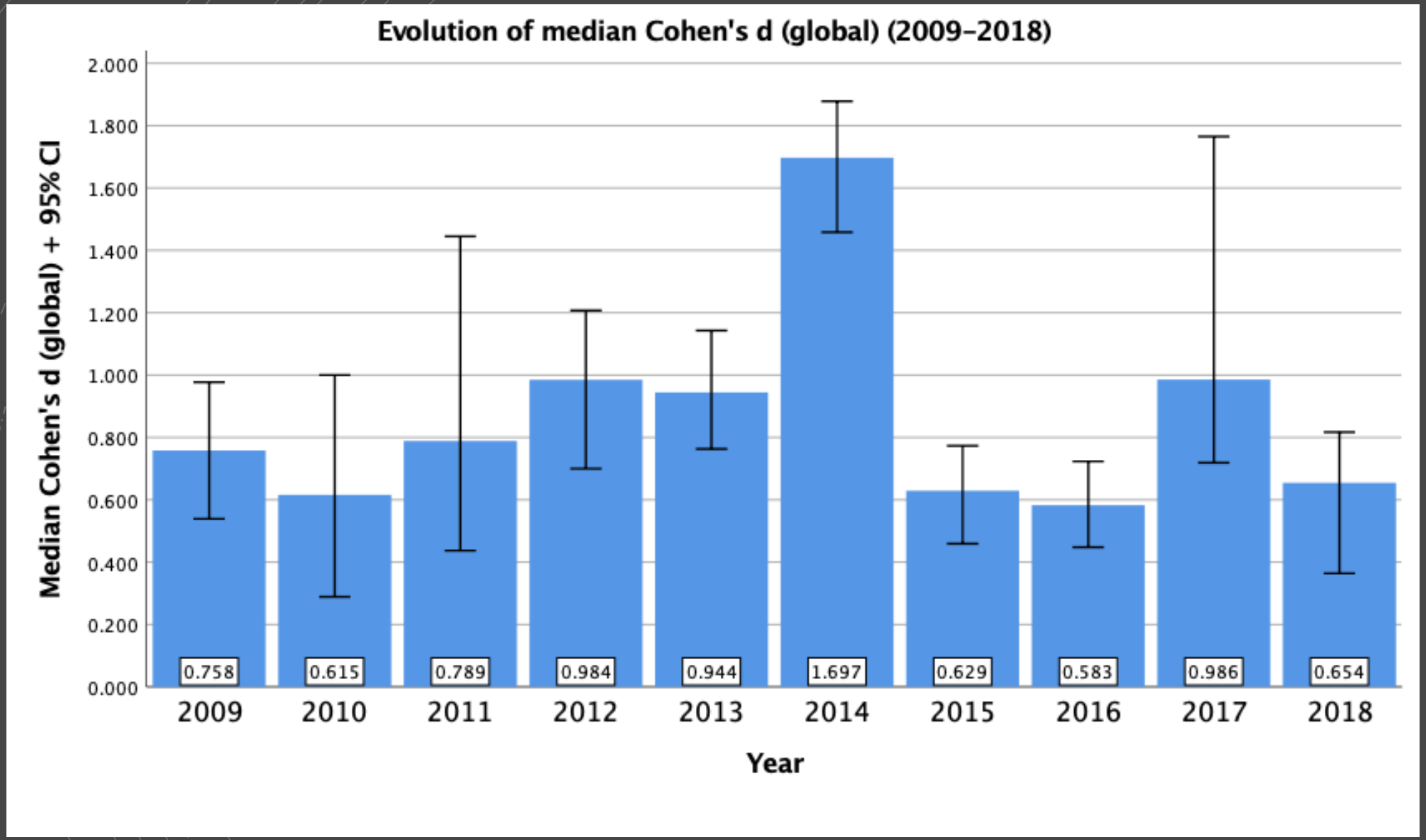

Estimated sample size for an independent t test; power = 0.80; α = 0.05
Sample size
Power
Effect size
Upper limit (1.317)
large d (1.150)
Lower limit (1.038)
Upper limit (0.742)
medium d (0.654)
Lower limit (0.568)
Upper limit (0.364)
small d (0.298)
Lower limit (0.219)
0
20
40
60
80
100
120
140
160
180
200
.05
.10
.15
.20
.25
.30
.35
.40
.45
.50
.55
.60
.65
.70
.75
.80
.85
.90
.95


Estimated sample size for a fixed effects one-way ANOVA with 3 groups; power = 0.80; α = 0.05
Sample size
200
180
160
140
120
100
80
60
40
20
0
.05
.10
.15
.20
.25
.30
.35
.40
.45
.50
.55
.60
.65
.70
.75
.80
.85
.90
.95
Power
Effect size
Upper limit (0,734)
large f (0,632)
Lower limit (0,562)
Upper limit (0,523)
medium f (0,420)
Lower limit (0,369)
Upper limit (0,347)
small f (0,274)
Lower limit (0,223)

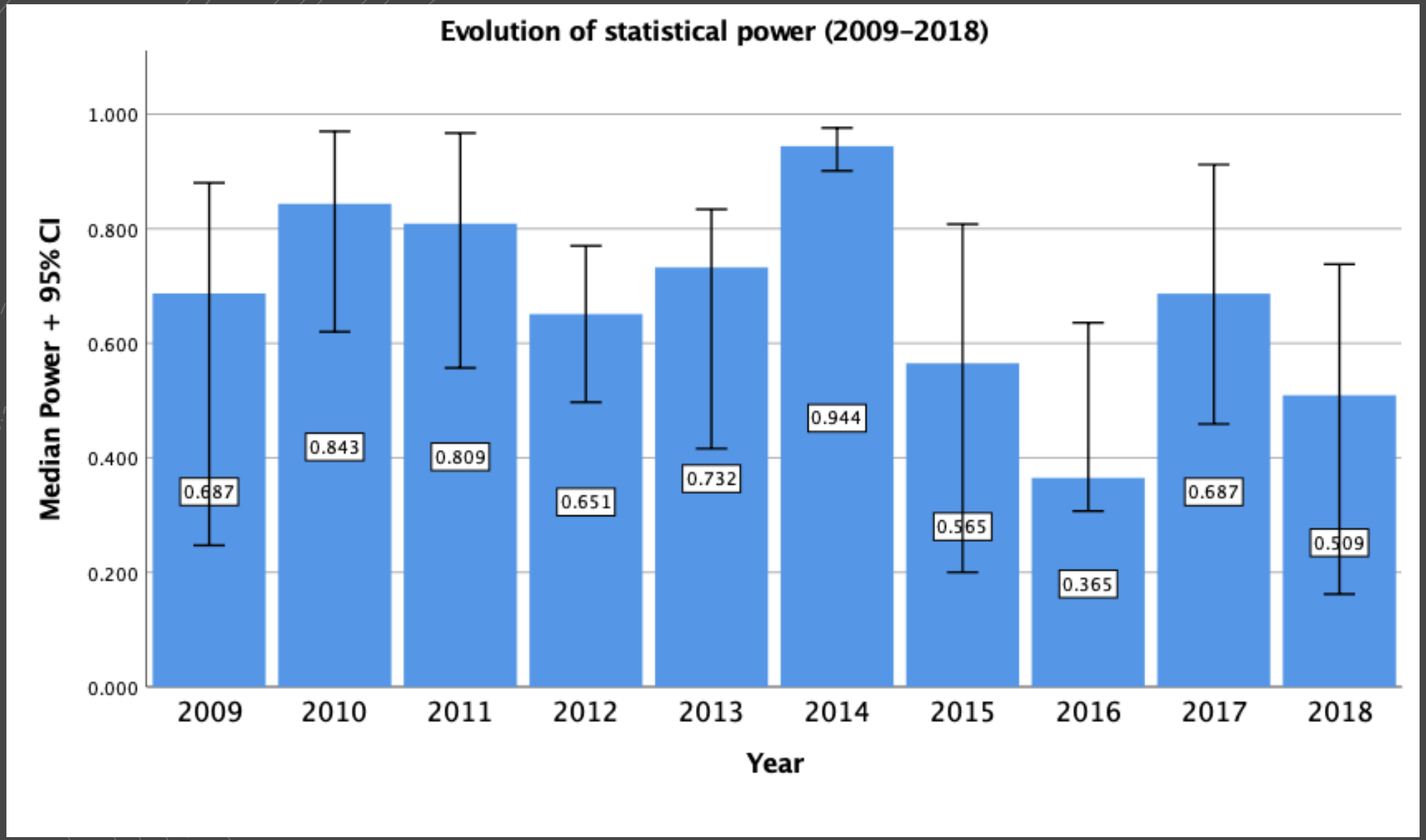

Estimated sample size for a correlation; power = 0.80; α = 0.05
Sample size
Power
Effect size
Upper limit (0,816)
large ρ (0,786)
Lower limit (0,756)
Upper limit (0,690)
medium ρ (0,648)
Lower limit (0,595)
Upper limit (0,513)
small ρ (0,470)
Lower limit (0,450)
0
10
20
30
40
50
60
.05
.10
.15
.20
.25
.30
.35
.40
.45
.50
.55
.60
.65
.70
.75
.80
.85
.90
.95

Evolution of sample size, effect size and power

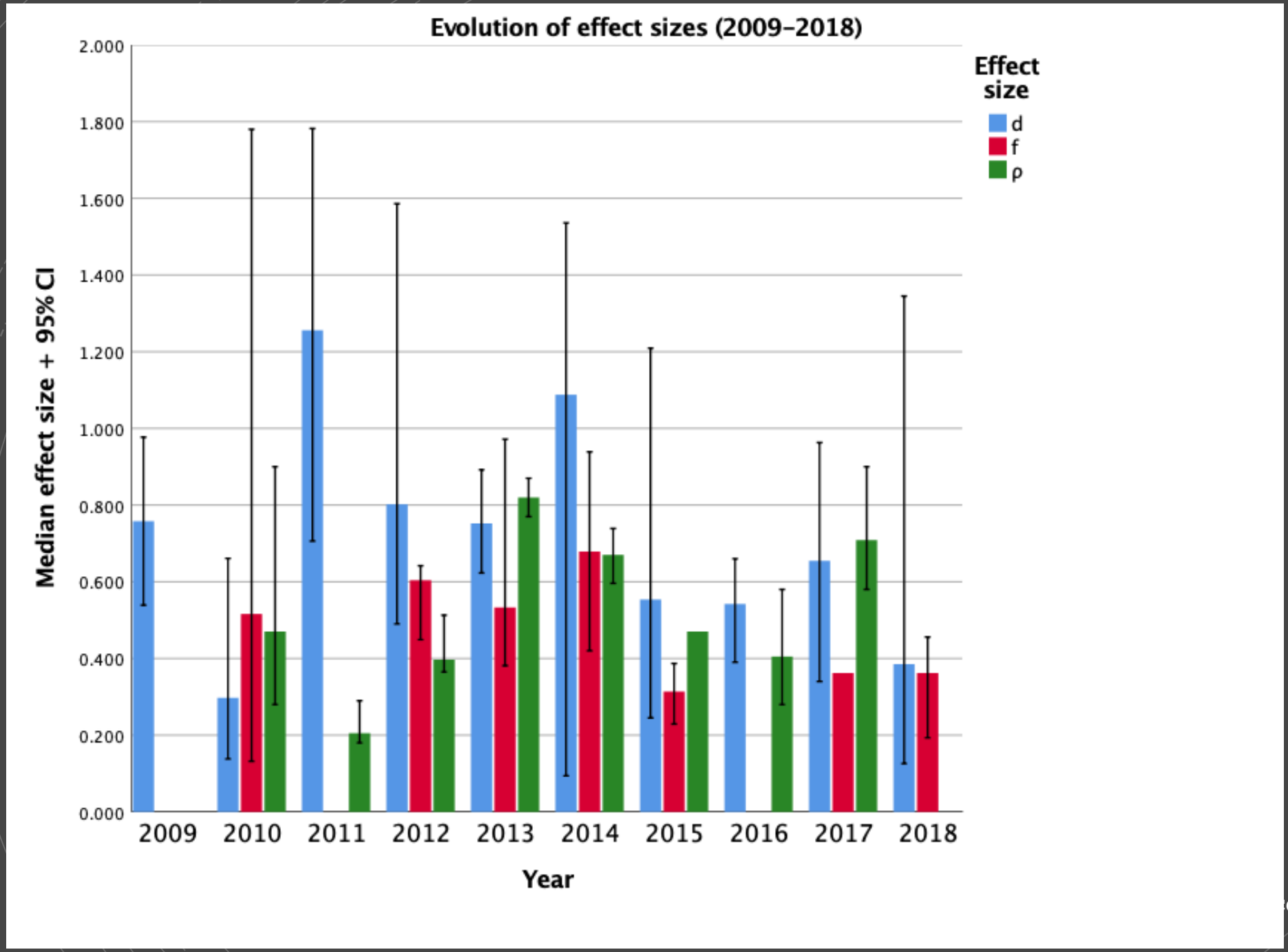

Evolution of median sample size (2009–2018)
Median sample size + 95% CI
160
140
120
100
80
60
40
20
0
15
18
14
20
16
30
154
24
19
68
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year



Cohen's *d* thresholds (Cohen, 1988): small = 0.2; medium = 0.5; large = 0.8



Recommended power (Cohen, 1988) = 0.80


Evolution of effect sizes (2009–2018)
Median effect size + 95% CI
2.000
1.800
1.600
1.400
1.200
1.000
0.800
0.600
0.400
0.200
0.000
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year
Effect size
d
f
ρ

Conclusions

- The median sample size is 21, the median effect size is over 0.80 (large) and power almost reaches 0.8 (0.7).
  - BUT: from 121 that included tests, in 46.2% it was not possible to extract effect sizes or even sample sizes.
  - BUT: selective reporting occurs VERY frequently.
  - HENCE: reality might not be as positive as we have observed.
- Non-significant tests tend to have lower sample sizes.
- Tests with effect size *d* tend to have lower power than ANOVAs and correlations.
- Sample size, effect size and power do not increase nor decrease as time advances.

Recommendations

- Fully report ALL statistical tests.
- Cooperate to increase sample size.
- Carry out sample size estimations.
- Determine relevant estimated effect sizes and significance levels (why always 0.05?)
- Replicate to test effect size magnitudes.

谢谢大家。

Christian.Olalla(a)uab.cat