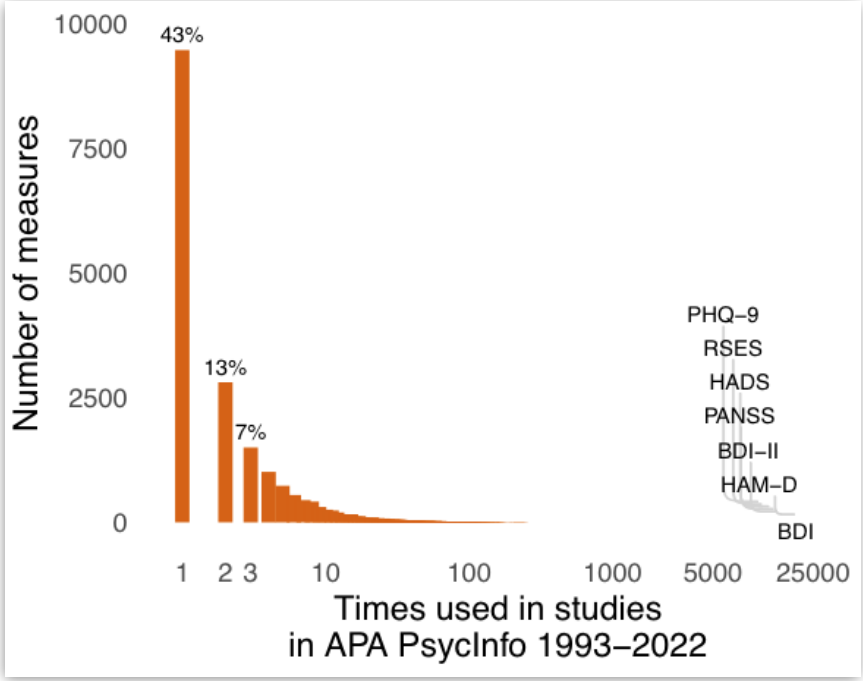
1
Measurement checks

Ad hoc measures preclude validation

- 43% of measures used just 1 time
- 80% used ⩽ 10 times

60,000 measures created since 1896

(Elson, Hussey, Alsalti, & Arslan, 2023)



Common measurement fallacies

A measure's name tells you what it measures

Nominal Fallacy

Common measurement fallacies

A measure's name tells you what it measures

Nominal Fallacy

Measures with the same name measure the same thing

Jingle Fallacy

Common measurement fallacies

A measure's name tells you what it measures

Nominal Fallacy

Measures with the same name measure the same thing

Jingle Fallacy

Measures with different names measure different things

Jangle Fallacy

What measurement fallacies may be at work here?

Correlations between different Theory of Mind tasks are near zero

Warnell & Redcay (2019)

Table 6
Relations among theory of mind tasks in adulthood.

	Spontaneous ToM	Belief Reasoning Speed	Pragmatics	Adult Eyes	Higher-Order ToM
Spontaneous ToM	–	−0.023	0.015	−0.115	0.125
Belief Reasoning Speed		–	0.056	0.115	0.048
Pragmatics			–	0.068	−0.051
Adult Eyes				–	−0.069
High-Order ToM					–

Note. Correlation values are Pearson's *r* controlling for age and gender.

Theory of Mind

What measurement fallacies may be at work here?

Measures of Implicit Attitudes / Bias (vs. explicit attitudes / self-report)

- Different implicit measures correlate poorly with one another ($r = .30$)
- Implicit and explicit measures correlate well with one-another ($r = .60$)

What measurement fallacies may be at work here?

Social Media is a Major Cause of the Mental Illness Epidemic in Teen Girls. Here's the Evidence.

Journalists should stop saying that the evidence is just correlational

JON HAIDT

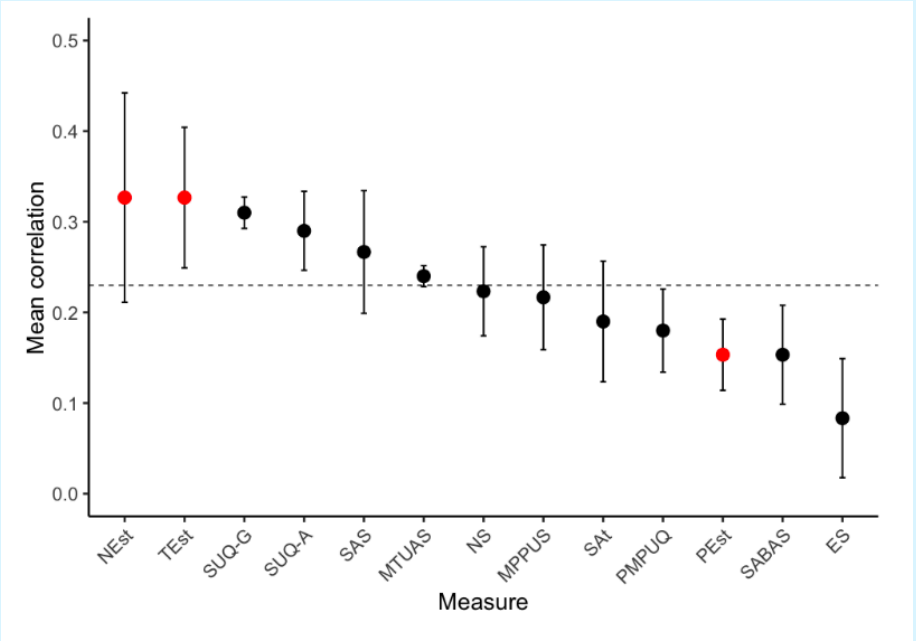
22 FEB 2023

Smartphone Use

What measurement fallacies may be at work here?

Self-reported smartphone usage
is weakly correlated
with actual usage

Ellis et al. 2019



Smartphone Use

What measurement fallacies may be at work here?

The Dot-Probe Task is used as a measure of attentional bias in anxiety

- Core foundation of experimental psychopathology

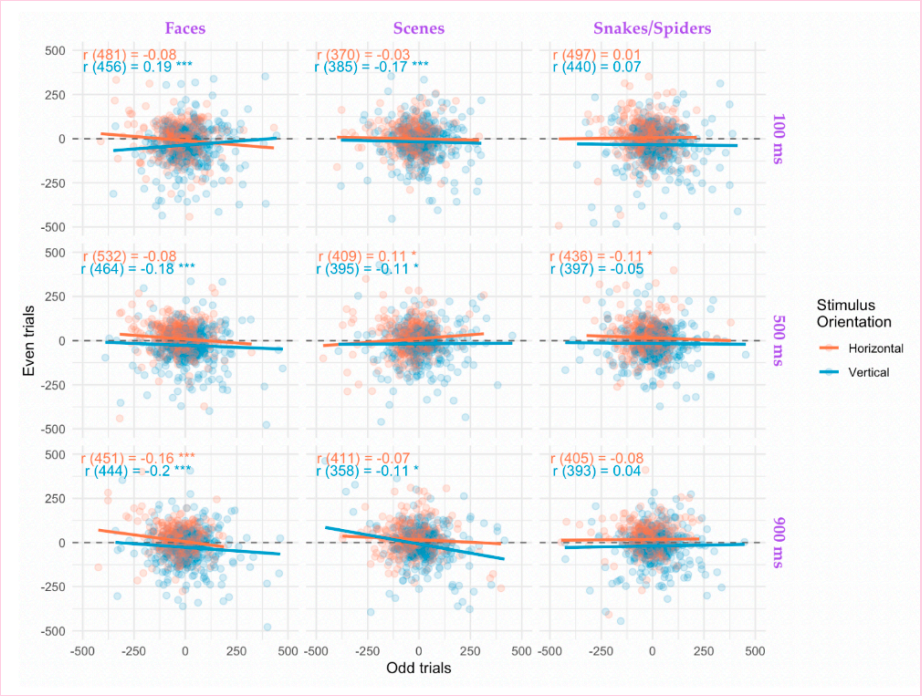
There is no one ‘Dot-Probe’ task, but many variants

- Stimuli domains
- Stimulus orientations
- Scoring comparisons
- Stimulus onset asynchrony

Dot-Probe Task

What measurement fallacies may be at work here?

Comparing 36 versions
in 9000 participants,
reliability was consistently near zero.
Xu et al. 2022



Dot-Probe Task

What measurement fallacies may be at work here?

In the field of Sex Research, many researchers wish to understand the context, causes, consequences, etc. of pornography use.

What could go wrong here, in terms of measurement validity?

[hint: think stupider]

Pornography use

What measurement fallacies may be at work here?

“57 of the 100 studies sampled
relied on idiosyncratic measures of pornography use
that were not found in other published research.”
Kohut et al. 2020

Issues with measuring pornography use

- What is “pornography”?
- What is “use”?
- What timeframe? (day, week, month, year, lifetime)
- What dimension? (frequency, first usage, specific content)

Pornography use

What measurement fallacies may be at work here?

“We controlled for socio-economic status”

SES

What measurement fallacies may be at work here?

“We controlled for socio-economic status”

There is no agreement how to measure SES
Antonoplis (2022)

Indicators and Modeling Procedures of Socioeconomic Status

Operationalizations	Percentage	Number (n=495)	Example
Indicators			
None given	0.20	1	–
Assets/Housing	8.10	40	Home value, Own home/car, Neighborhood wealth/cohesion
Composite	5.25	26	Hollingshead Index, Brazilian ABEP
Demographic	0.40	2	Race, Gender
Education	27.47	136	Personal education (highest degree attained)
Family structure	4.24	21	Teen mom, Father present, Number of children
Income/Poverty	22.83	113	Family income, Neighborhood poverty rate
Mannerisms	2.83	14	Extracurricular activities, Verb use, Name, Clothing
Occupation	16.36	81	Parental Duncan’s SEI
Subjective	10.30	51	MacArthur ladder
Uncategorized	2.02	10	Health insurance, Welfare/aid, Food insecurity

SES

Exercise: Check your assigned articles

- Are the measures *ad hoc* or established?

- What validity evidence do they present against jingle/jangle?

2
Magnitude checks

We need better intuitions for effect size plausibility

“Chocolates are more desirable than poop”

Balcetis & Dunning (2010) Study 3b pretest

We need better intuitions for effect size plausibility

"Chocolates are more desirable than poop"

Cohen's d = 4.5, 95% CI [3.3, 5.7]

Balcetis & Dunning (2010) Study 3b pretest

We need better intuitions for effect size plausibility

"Chocolates are more desirable than poop"

Cohen's d = 4.5, 95% CI [3.3, 5.7]

Balcetis & Dunning (2010) Study 3b pretest

Asking children "Boys tend to wear skirts" vs. "Girls tend to wear skirts"

Streck & Kessels (2024)

We need better intuitions for effect size plausibility

"Chocolates are more desirable than poop"

Cohen's d = 4.5, 95% CI [3.3, 5.7]

Balcetis & Dunning (2010) Study 3b pretest

Asking children "Boys tend to wear skirts" vs. "Girls tend to wear skirts"

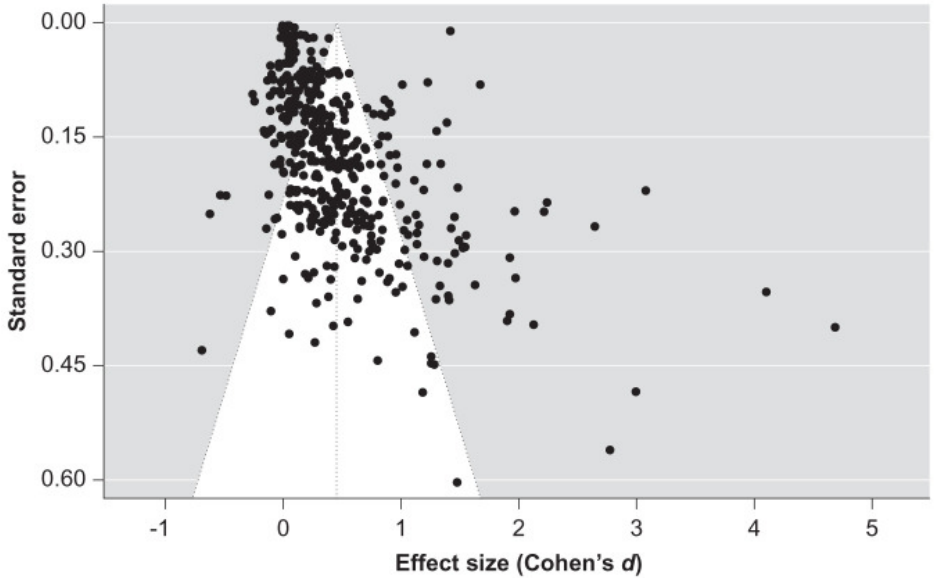
Cohen's d = 5.5, 95% CI [4.9, 6.1]

Streck & Kessels (2024)

Cohen's *d* > 2 raise serious doubts

Cohen's *d* > 5 are silly*

*in normal circumstances



Mertens et al. (2021) The effectiveness of nudging

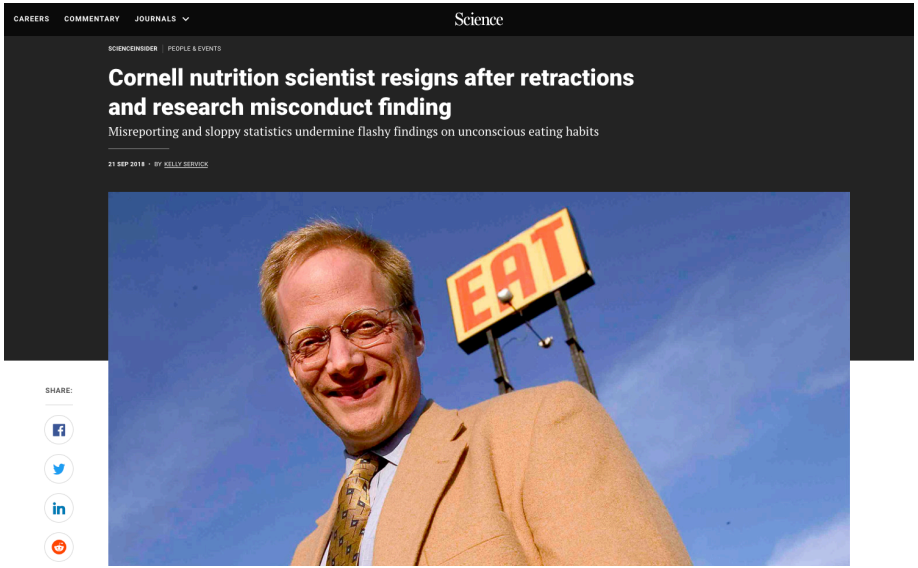
83. *A Geier, B Wansink, P Rozin, Red potato chips: Segmentation cues can substantially decrease food intake. *Heal. Psychol.* **31**, 398–401 (2012).

191. *M Shimizu, CR Payne, B Wansink, When snacks become meals: How hunger and environmental cues bias food intake. *Int. J. Behav. Nutr. Phys. Activity* **7**, 1–6 (2010).

217. *E van Kleef, M Shimizu, B Wansink, Serving bowl selection biases the amount of food served. *J. Nutr. Educ. Behav.* **44**, 66–70 (2012).

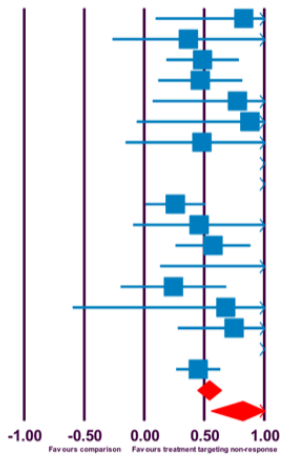
220. *B Wansink, AS Hanks, Slim by design: Serving healthy foods first in buffet lines improves overall meal selection. *PLoS ONE* **8**, 1–5 (2013).
221. *B Wansink, J Kim, Bad popcorn in big buckets: Portion size can influence intake as much as taste. *J. Nutr. Educ. Behav.* **37**, 242–245 (2005).
222. *B Wansink, K van Ittersum, Bottoms up! the influence of elongation on pouring and consumption volume. *J. Consumer Res.* **30**, 455–463 (2003).
223. *B Wansink, A Cardello, J North, Fluid consumption and the potential role of canteen shape in minimizing dehydration. *Mil. Medicine* **170**, 871–873 (2005).
224. *B Wansink, JE Painter, J North, Bottomless bowls: Why visual cues of portion size may influence intake. *Obes. Res.* **13**, 93–100 (2005).
225. *B Wansink, D Soman, KC Herbst, Larger partitions lead to larger sales: Divided grocery carts alter purchase norms and increase sales. *J. Bus. Res.* **75**, 202–209 (2017).

226. *B Wansink, K van Ittersum, JE Painter, Ice cream illusions: Bowls, spoons, and self-served portion sizes. *Am. J. Prev. Medicine* **31**, 240–243 (2006).
227. *B Wansink, K van Ittersum, CR Payne, Larger bowl size increases the amount of cereal children request, consume, and waste. *The J. Pediatr.* **164**, 323–326 (2014).



Mertens et al. (2021) The effectiveness of nudging

Model	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	Barnhofer et al. (2009)	0.829	0.374	0.140	0.095	1.563	2.213	0.027
	Clarke et al. (2014)	0.368	0.323	0.104	-0.265	1.001	1.140	0.254
	Eisendrath et al. (2016)	0.486	0.154	0.024	0.184	0.789	3.151	0.002
	Fonagy et al. (2015)	0.467	0.179	0.032	0.117	0.817	2.612	0.009
	Gloster et al. (2015)	0.777	0.361	0.130	0.069	1.484	2.152	0.031
	Harley et al. (2008)	0.881	0.481	0.232	-0.062	1.824	1.831	0.067
	Hartmann Souza et al. (2016)	0.480	0.324	0.105	-0.155	1.116	1.481	0.139
	Hinton et al. (2005)	2.226	0.402	0.162	1.438	3.015	5.532	0.000
	Isasi et al. (2010)	5.233	0.665	0.442	3.929	6.536	7.868	0.000
	Kocsis et al. (2009)	0.259	0.126	0.016	0.012	0.506	2.055	0.040
	Ludman et al. (2007)	0.458	0.281	0.079	-0.093	1.008	1.629	0.103
	Mantani et al. (2017)	0.572	0.159	0.025	0.260	0.885	3.592	0.000
	Moore & Blackburn (1997)	1.653	0.776	0.602	0.132	3.173	2.130	0.033
	Nakagawa et al. (2017)	0.243	0.224	0.050	-0.197	0.683	1.084	0.278
	Otto et al. (2003)	0.679	0.650	0.423	-0.596	1.954	1.044	0.297
	Town et al. (2017)	0.748	0.239	0.057	0.280	1.216	3.131	0.002
	Watanabe et al. (2011)	1.811	0.391	0.153	1.044	2.578	4.627	0.000
	Wiles et al. (2013)	0.450	0.094	0.009	0.266	0.633	4.807	0.000
Fixed		0.540	0.050	0.003	0.442	0.638	10.765	0.000
Random		0.818	0.134	0.018	0.556	1.081	6.106	0.000



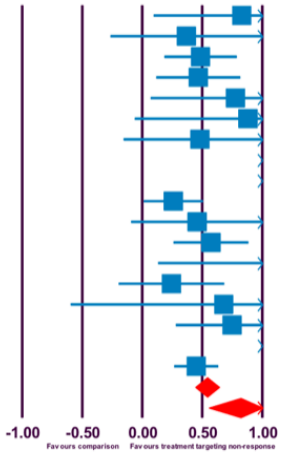
Effectiveness at post-treatment on symptom reduction

Gloster et al. (2020) Treating treatment non-responders

Model	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	Barnhofer et al. (2009)	0.829	0.374	0.140	0.095	1.563	2.213	0.027
	Clarke et al. (2014)	0.368	0.323	0.104	-0.265	1.001	1.140	0.254
	Eisendrath et al. (2016)	0.486	0.154	0.024	0.184	0.789	3.151	0.002
	Fonagy et al. (2015)	0.467	0.179	0.032	0.117	0.817	2.612	0.009
	Gloster et al. (2015)	0.777	0.361	0.130	0.069	1.484	2.152	0.031
	Harley et al. (2008)	0.881	0.481	0.232	-0.062	1.824	1.831	0.067
	Hartmann Souza et al. (2016)	0.480	0.324	0.105	-0.155	1.116	1.481	0.139
	Hinton et al. (2005)	2.226	0.402	0.162	1.438	3.015	5.532	0.000
	Isasi et al. (2010)	5.233	0.665	0.442	3.929	6.536	7.868	0.000
	Kocsis et al. (2009)	0.259	0.126	0.016	0.012	0.506	2.055	0.040
	Ludman et al. (2007)	0.458	0.281	0.079	-0.093	1.008	1.629	0.103
	Mantani et al. (2017)	0.572	0.159	0.025	0.260	0.885	3.592	0.000
	Moore & Blackburn (1997)	1.653	0.776	0.602	0.132	3.173	2.130	0.033
	Nakagawa et al. (2017)	0.243	0.224	0.050	-0.197	0.683	1.084	0.278
	Otto et al. (2003)	0.679	0.650	0.423	-0.596	1.954	1.044	0.297
	Town et al. (2017)	0.748	0.239	0.057	0.280	1.216	3.131	0.002
	Watanabe et al. (2011)	1.811	0.391	0.153	1.044	2.578	4.627	0.000
	Wiles et al. (2013)	0.450	0.094	0.009	0.266	0.633	4.807	0.000
Fixed		0.540	0.050	0.003	0.442	0.638	10.765	0.000
Random		0.818	0.134	0.018	0.556	1.081	6.106	0.000

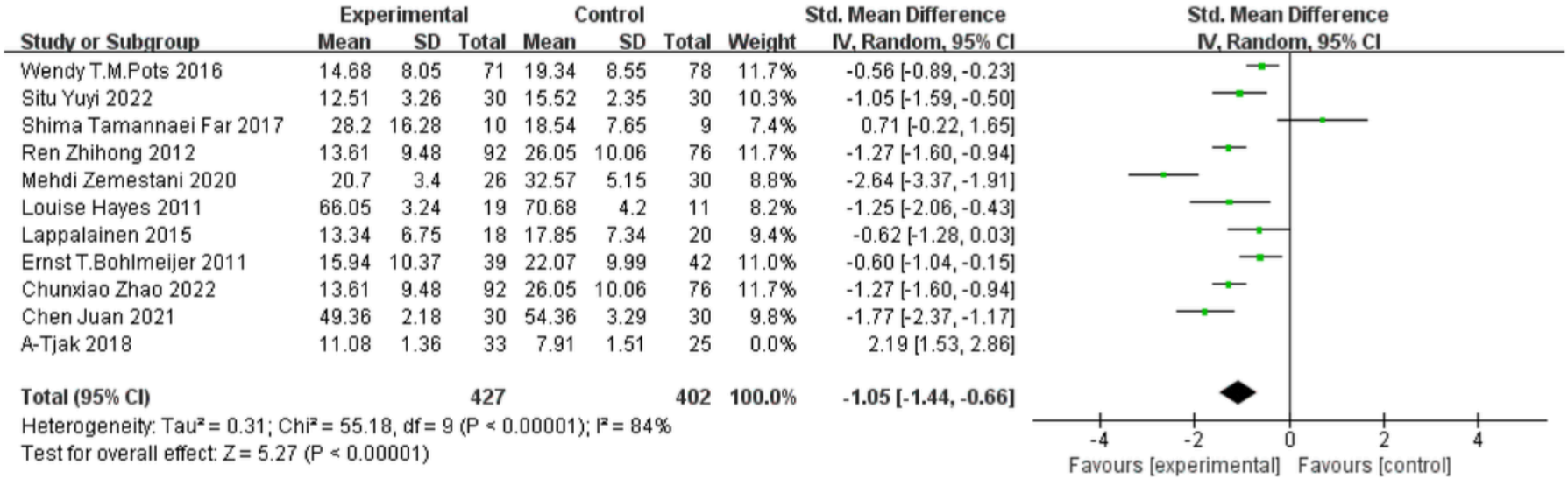
Statistics for each study

Std diff in means and 95%CI



Effectiveness at post-treatment on symptom reduction

Gloster et al. (2020) Treating treatment non-responders

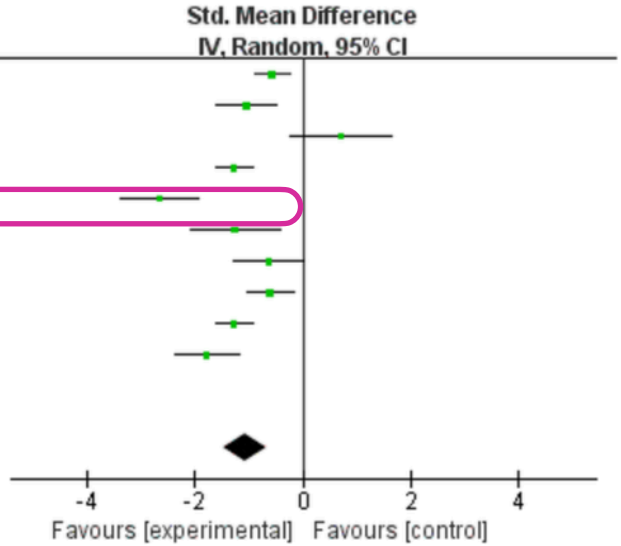


Zhao et al. (2023) Effect of acceptance and commitment therapy for depressive disorders

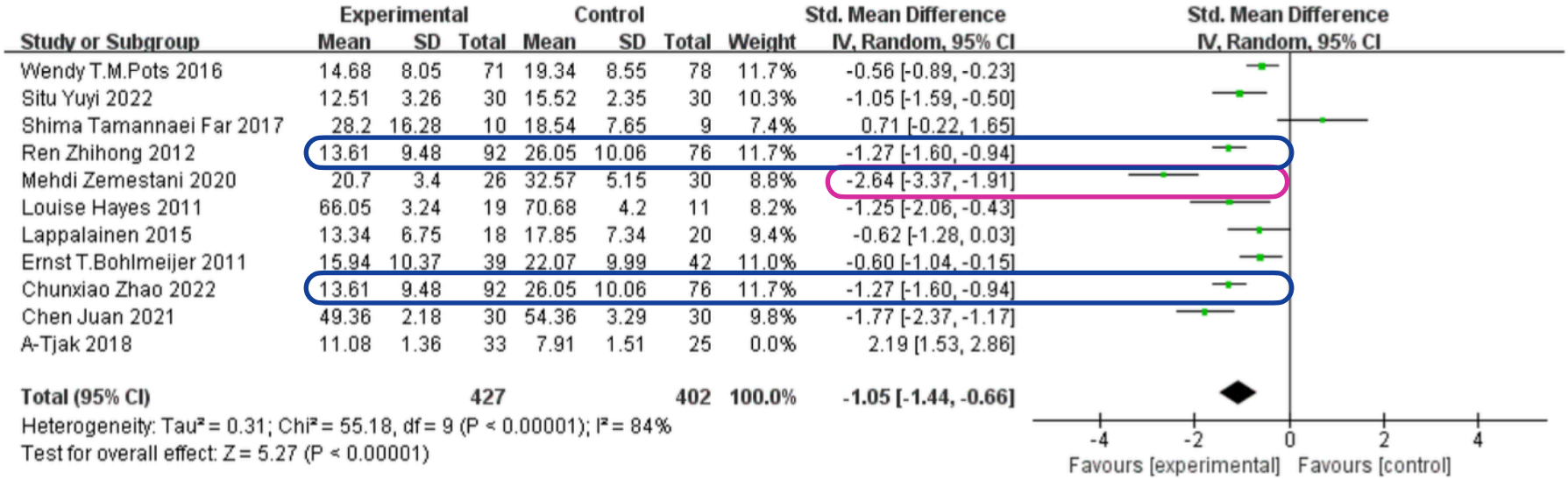
Study or Subgroup	Experimental Mean	SD	Total	Control Mean	SD	Total	Weight	Std. Mean Difference IV, Random, 95% CI
Wendy T.M.Pots 2016	14.68	8.05	71	19.34	8.55	78	11.7%	-0.56 [-0.89, -0.23]
Situ Yuyi 2022	12.51	3.26	30	15.52	2.35	30	10.3%	-1.05 [-1.59, -0.50]
Shima Tamannaei Far 2017	28.2	16.28	10	18.54	7.65	9	7.4%	0.71 [-0.22, 1.65]
Ren Zhihong 2012	13.61	9.48	92	26.05	10.06	76	11.7%	-1.27 [-1.60, -0.94]
Mehdi Zemestani 2020	20.7	3.4	26	32.57	5.15	30	8.8%	-2.64 [-3.37, -1.91]
Louise Hayes 2011	66.05	3.24	19	70.68	4.2	11	8.2%	-1.25 [-2.06, -0.43]
Lappalainen 2015	13.34	6.75	18	17.85	7.34	20	9.4%	-0.62 [-1.28, 0.03]
Ernst T.Bohlmeijer 2011	15.94	10.37	39	22.07	9.99	42	11.0%	-0.60 [-1.04, -0.15]
Chunxiao Zhao 2022	13.61	9.48	92	26.05	10.06	76	11.7%	-1.27 [-1.60, -0.94]
Chen Juan 2021	49.36	2.18	30	54.36	3.29	30	9.8%	-1.77 [-2.37, -1.17]
A-Tjak 2018	11.08	1.36	33	7.91	1.51	25	0.0%	2.19 [1.53, 2.86]
Total (95% CI)			**427**			**402**	**100.0%**	**-1.05 [-1.44, -0.66]**

Heterogeneity: Tau² = 0.31; Chi² = 55.18, df = 9 (P < 0.00001); I² = 84%

Test for overall effect: Z = 5.27 (P < 0.00001)



Zhao et al. (2023) Effect of acceptance and commitment therapy for depressive disorders



Zhao et al. (2023) Effect of acceptance and commitment therapy for depressive disorders

Exercise: Check your assigned articles

- Recalculate effect sizes from summary statistics

- Check no-one confuses SE and SD
  - The “Standard Error error”
  - ~30% of meta-analyses do this
  - Maassen et al. (2023)