Vahey et al. 2015

Individual critique using Ilias submission

“Information that is repeated or overlooked”

Reported meta-analysis result

$\bar{r}$ = .45, 95% CI [.40, .54], 95% CR [.23, .67]

Reported meta-analysis result

$\bar{r}$ = .45, 95% CI [.40, .54], 95% CR [.23, .67]

-.05, .09 -.22, .22

Reported meta-analysis result

$\bar{r}$ = .45, 95% CI [.40, .54], 95% CR [.23, .67]

-.05, .09 -.22, .22

- **Asymmetric confidence intervals**
  - Often a sign something is wrong
  - Wider on left:
    - Maybe a transformation was used
    - Maybe an error
  - Wider on right: <u>very implausible</u>

Definitions and math

- **Check if definitions or math are wrong**

Vahey et al. (2015) define Credibility Intervals incorrectly

Gloster et al. (2020) define Cohen's d incorrectly:

surement tools. The standardized mean difference is calculated by the following equation:

SMD

= Difference in mean outcome between groups

/Change of SD of outcome between groups.

Reported power analyses

			Vahey et al. (2015)	
Test	Tails	Estimated using*	$\bar{r}$	N
Pearson's r	One	Point estimate	0.45	29
Pearson's r	One	Lower bound of 95% CI	0.40	37
Pearson's r	Two	Point estimate	0.45	36
Pearson's r	Two	Lower bound of 95% CI	0.40	-
Independent t-test (Cohen's d)**	One	Point estimate	1.01	26
Independent t-test (Cohen's d)**	One	Lower bound of 95% CI	0.87	36
Dependent t-test (Cohen's d) **	One	Point estimate	1.01	8
Dependent t-test (Cohen's d) **	One	Lower bound of 95% CI	0.87	10

Reported power analyses

			Vahey et al. (2015)		Verified
Test	Tails	Estimated using*	$\bar{r}$	N	N
Pearson's r	One	Point estimate	0.45	29	29
Pearson's r	One	Lower bound of 95% CI	0.40	37	37
Pearson's r	Two	Point estimate	0.45	36	36
Pearson's r	Two	Lower bound of 95% CI	0.40	-	46
Independent t-test (Cohen's d)**	One	Point estimate	1.01	26	26
Independent t-test (Cohen's d)**	One	Lower bound of 95% CI	0.87	36	34***
Dependent t-test (Cohen's d) **	One	Point estimate	1.01	8	8
Dependent t-test (Cohen's d) **	One	Lower bound of 95% CI	0.87	10	10

Reproduce results: not all reproducible

Reported power analyses

			Vahey et al. (2015)		Verified
Test	Tails	Estimated using*	$\bar{r}$	N	N
Pearson's r	One	Point estimate	0.45	29	29
Pearson's r	One	Lower bound of 95% CI	0.40	37	37
Pearson's r	Two	Point estimate	0.45	36	36
Pearson's r	Two	Lower bound of 95% CI	0.40	-	46
Independent t-test (Cohen's d)**	One	Point estimate	1.01	26	26
Independent t-test (Cohen's d)**	One	Lower bound of 95% CI	0.87	36	34***
Dependent t-test (Cohen's d) **	One	Point estimate	1.01	8	8
Dependent t-test (Cohen's d) **	One	Lower bound of 95% CI	0.87	10	10

Reproduce results: not all reproducible

Assess how liberal the test is: extremely generous

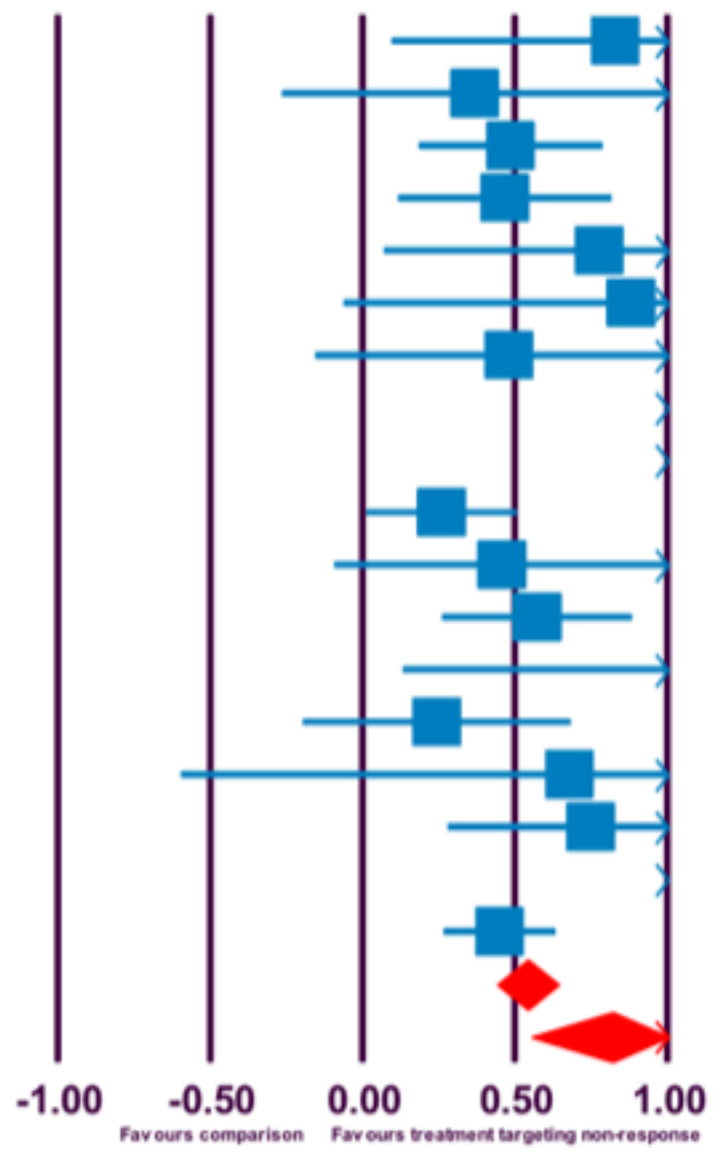
Check for extremely large effect sizes

Cohen's d > 2 are suspicious

Cohen's d > 5 are definitely wrong

Model	Study name	Statistics for each study: 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

Std diff in means and 95%CI



Effectiveness at post-treatment on symptom reduction

Gloster et al. (2020)

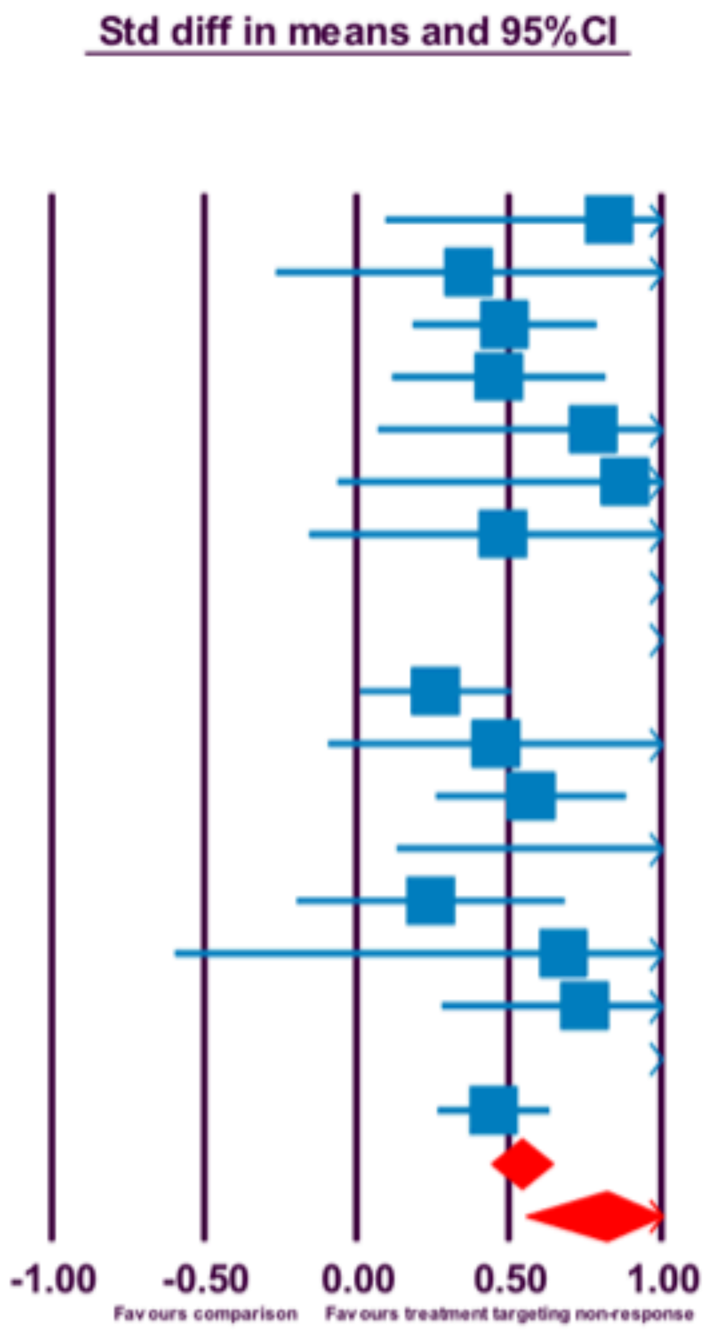
Check for extremely large effect sizes

Cohen's d > 2 are suspicious

Cohen's d > 5 are almost definitely wrong

Preference for chocolate over is Cohen's d = 4.5

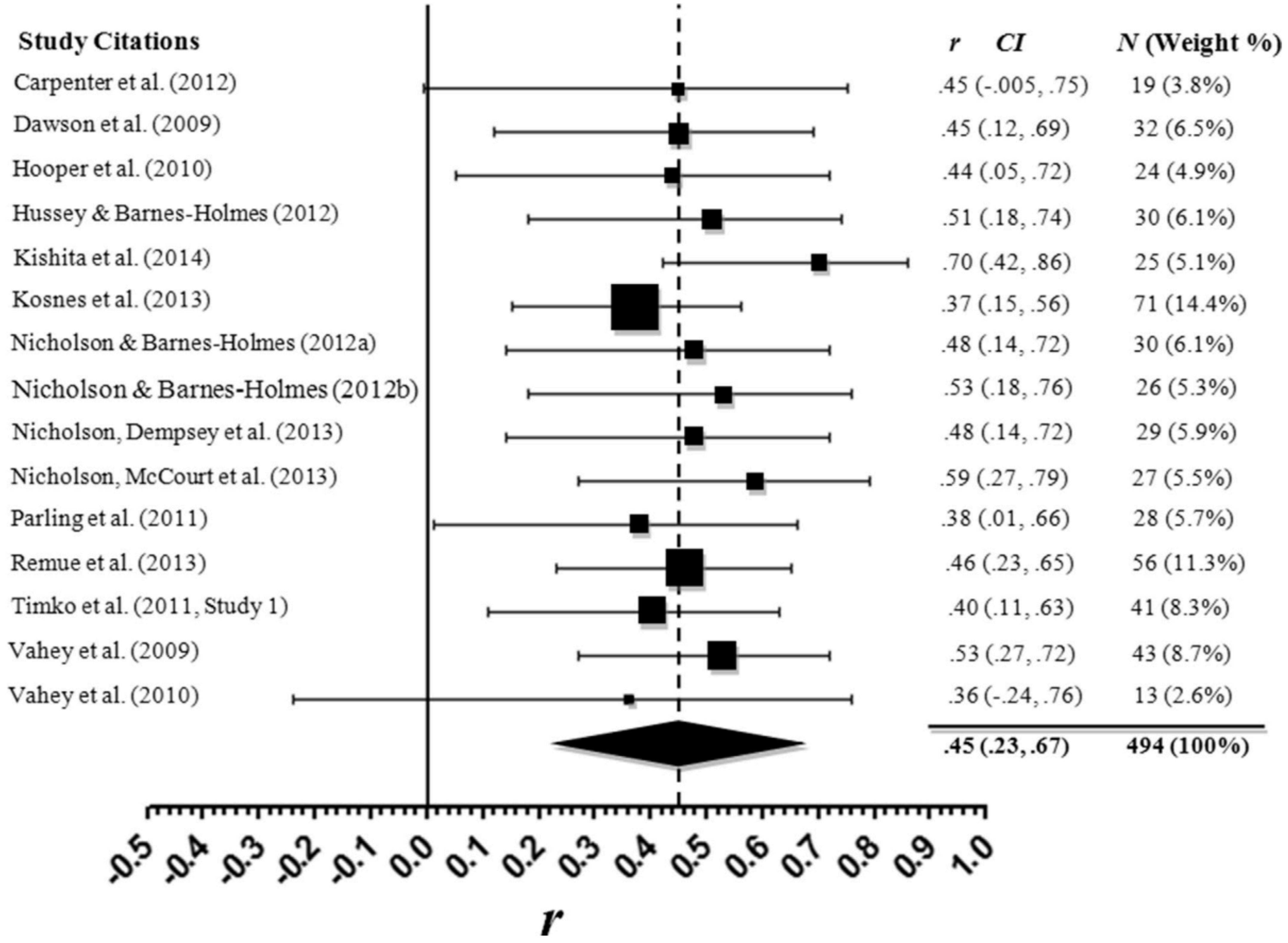
Model	Study name	Statistics for each study						
		Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	Barnhofer et al. (2009)	0.829	0.374	0.140	0.095	1.563	2.213	0.027
	Clarke et al. (2014)	0.368	0.323	0.104	-0.265	1.001	1.140	0.254
	Eisendrath et al. (2016)	0.486	0.154	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)

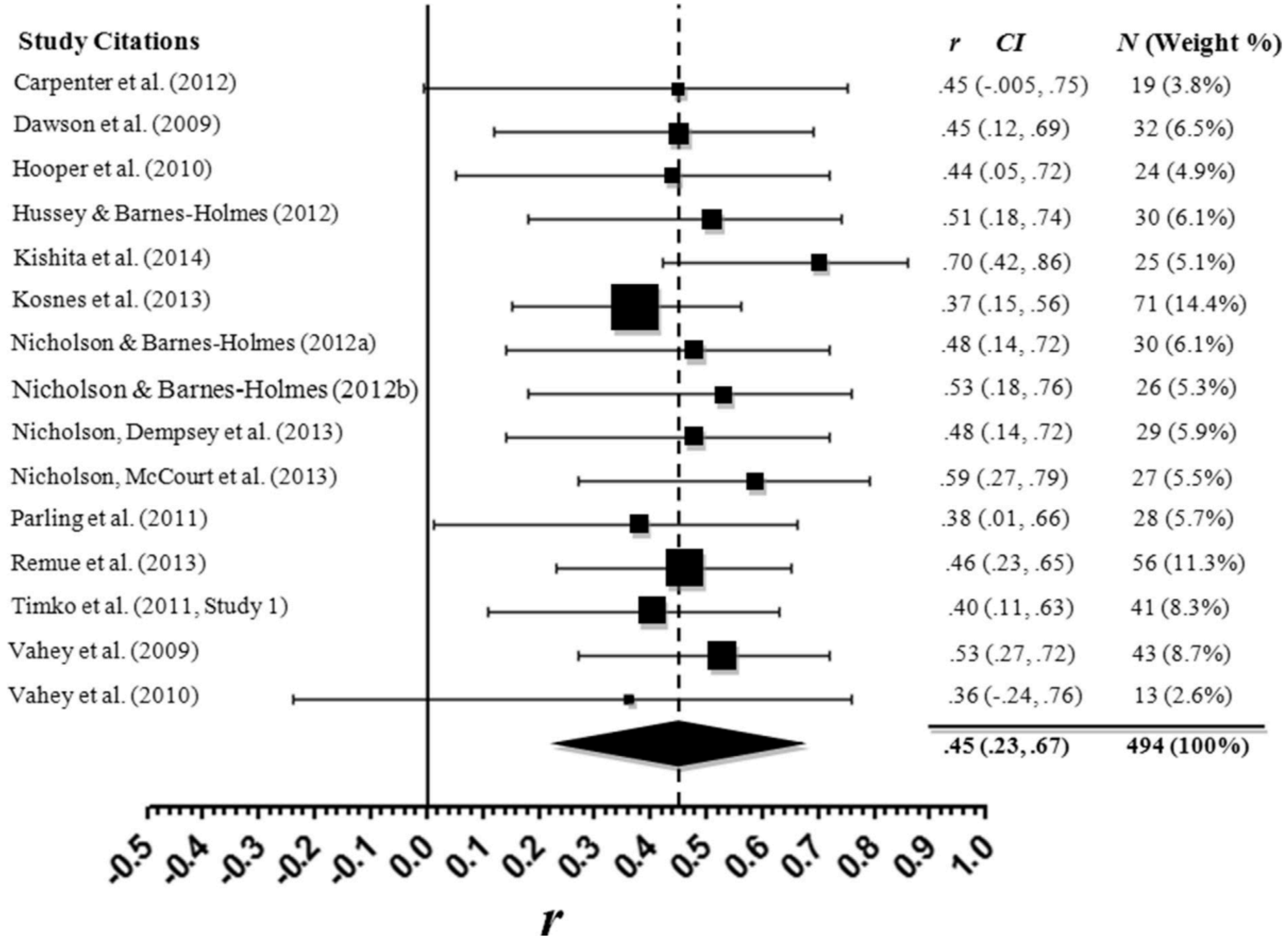
Reported forest plot and meta-analysis reported in text



Extracted from text:

$\bar{r}$ = .45, 95% CI [.40, .54], 95% CR [.23, .67]

Reported forest plot and meta-analysis reported in text



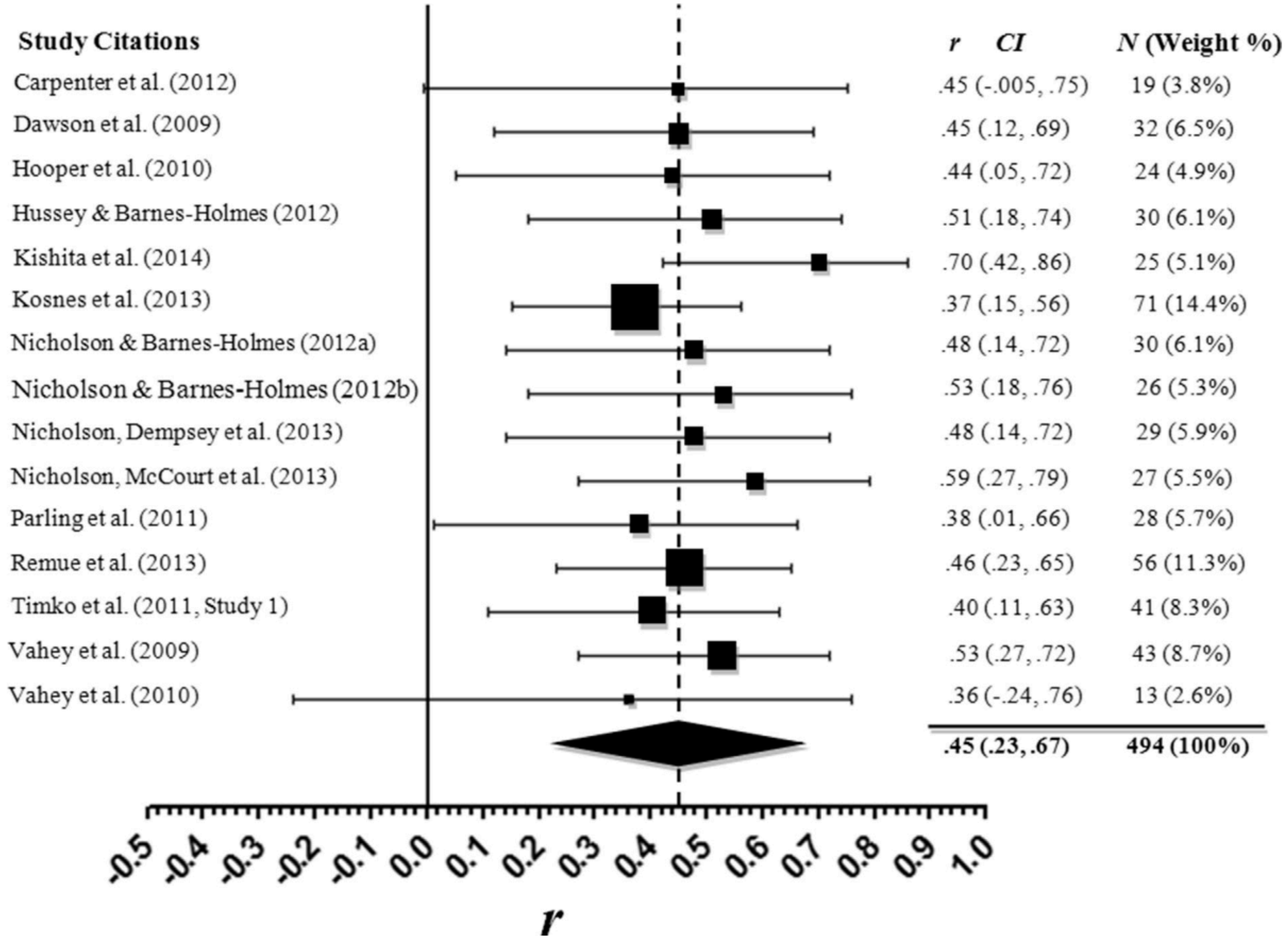
Extracted from text:

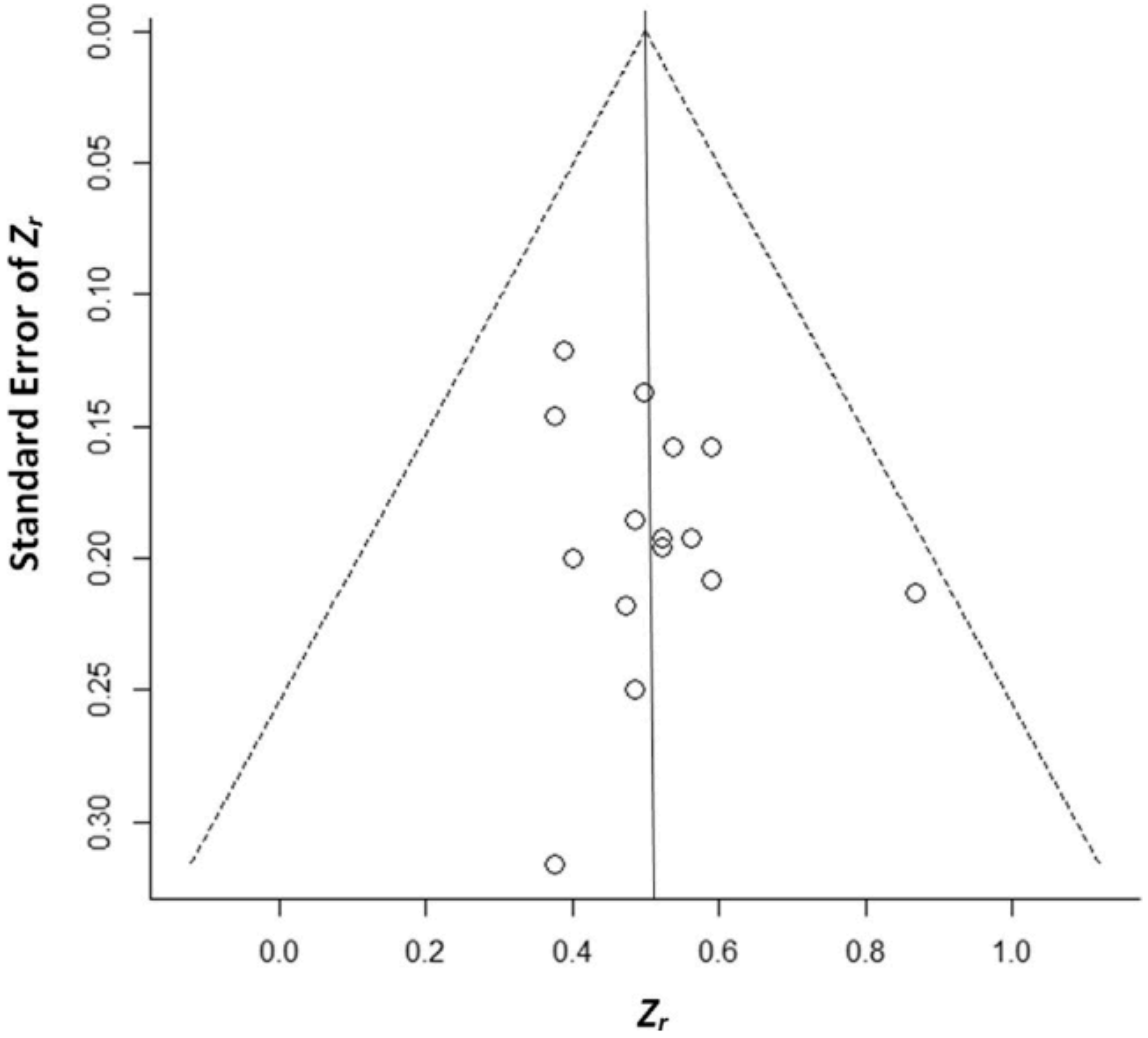
$\bar{r}$ = .45, 95% CI [.40, .54], 95% CR [.23, .67]

"Information that is repeated or overlooked"

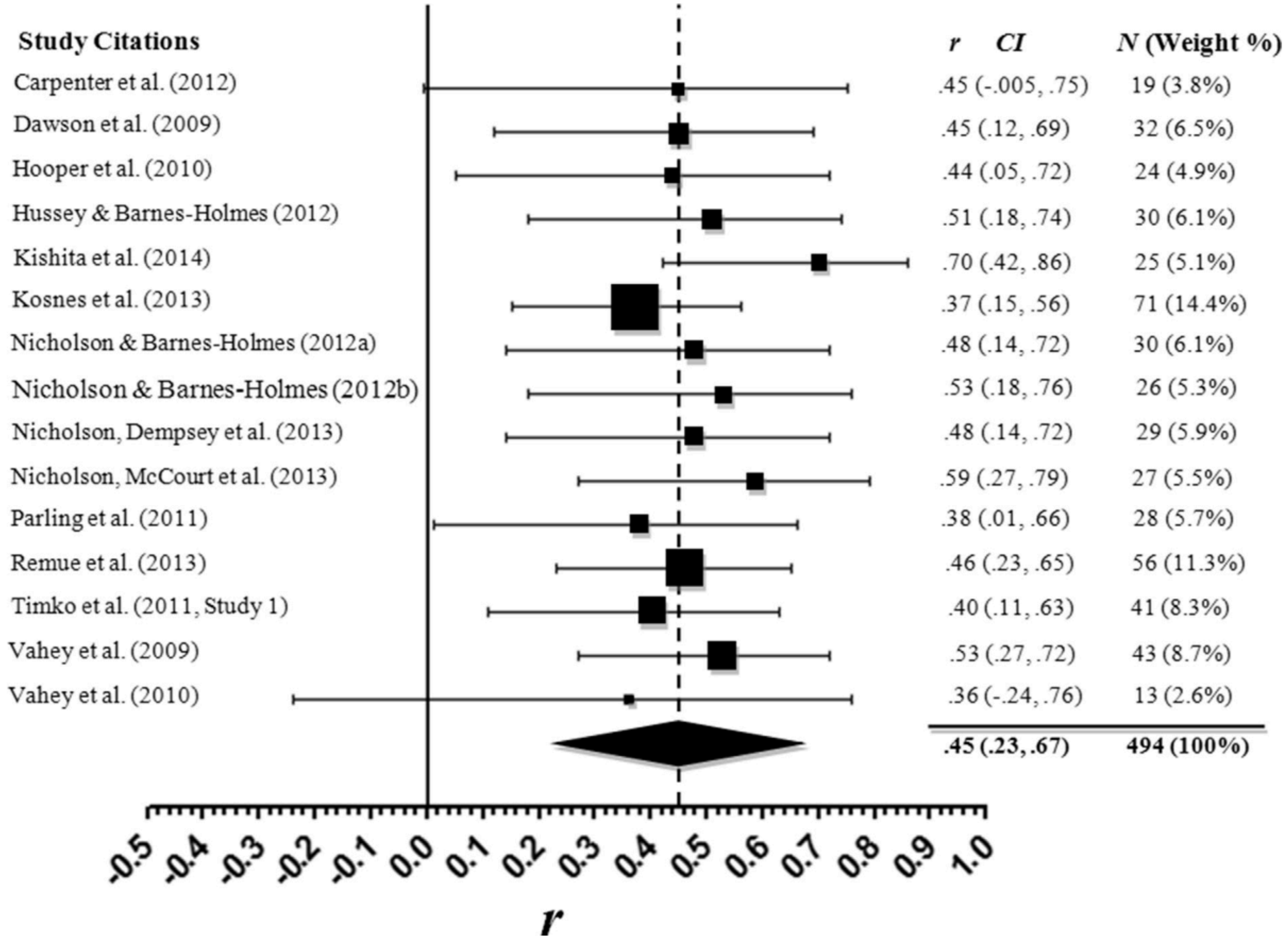
Compare text and plot: they match

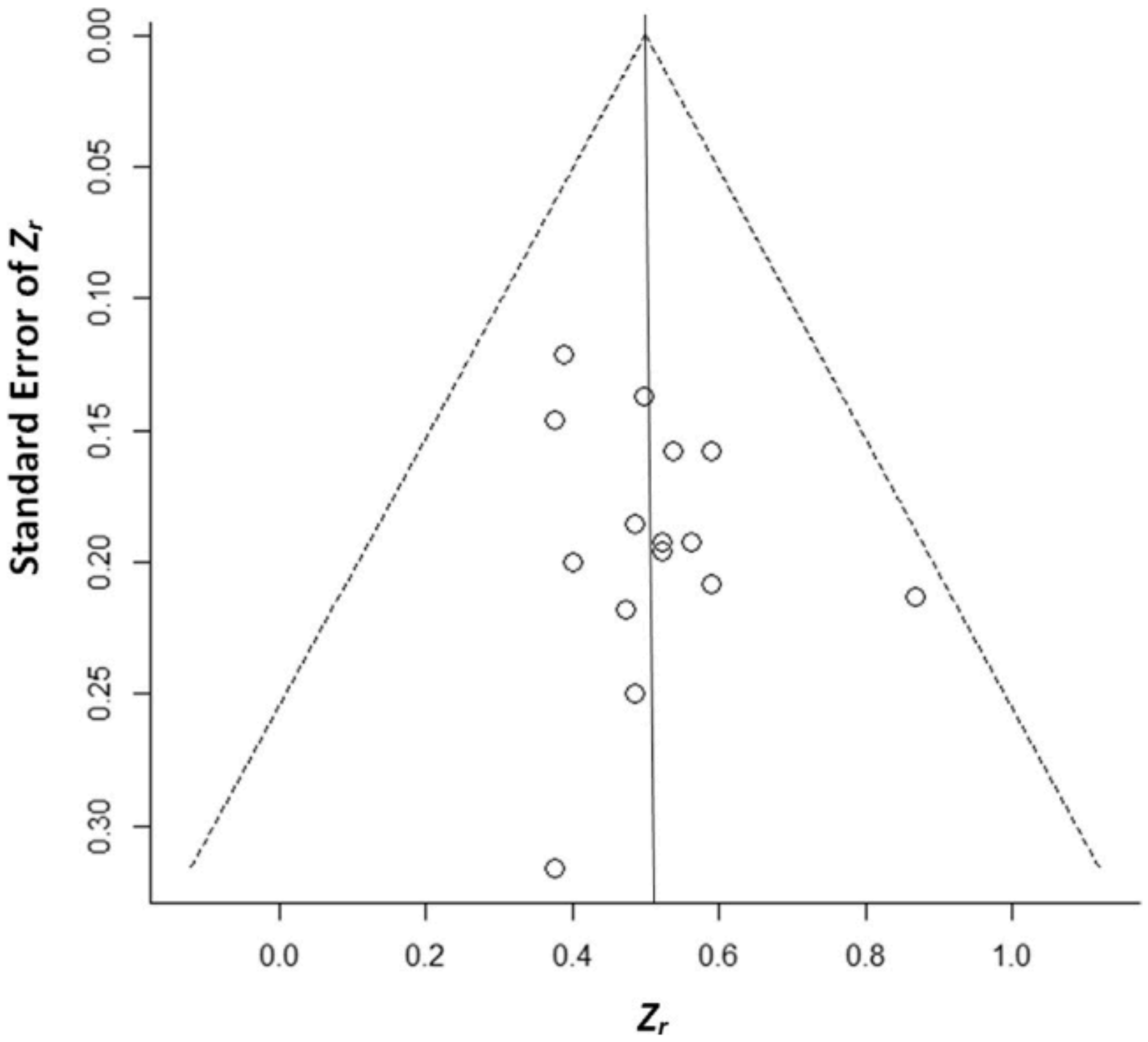
Reported forest plot and funnel plot





Reported forest plot and funnel plot

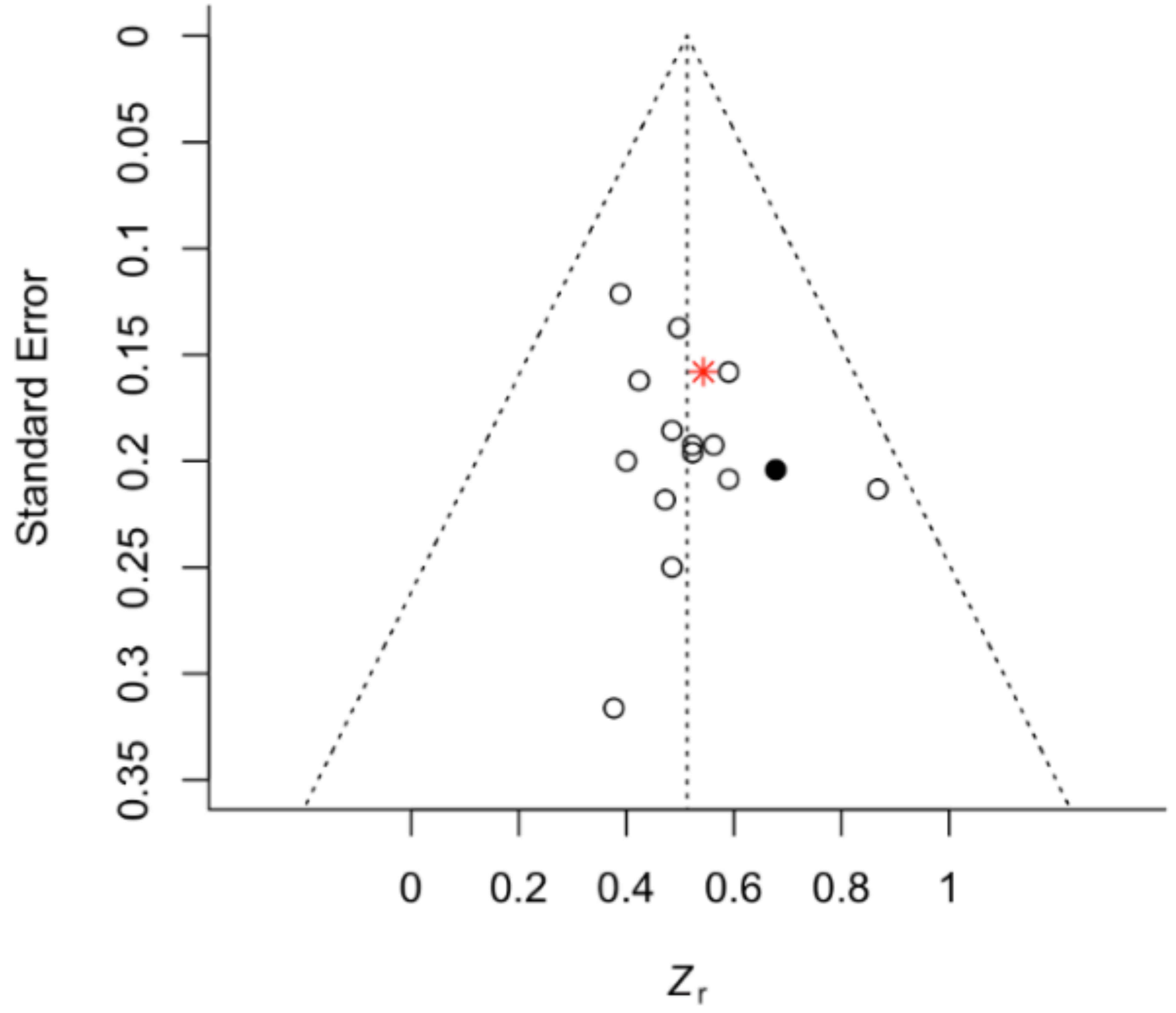




"Information that is repeated or overlooked"

Compare data in plots

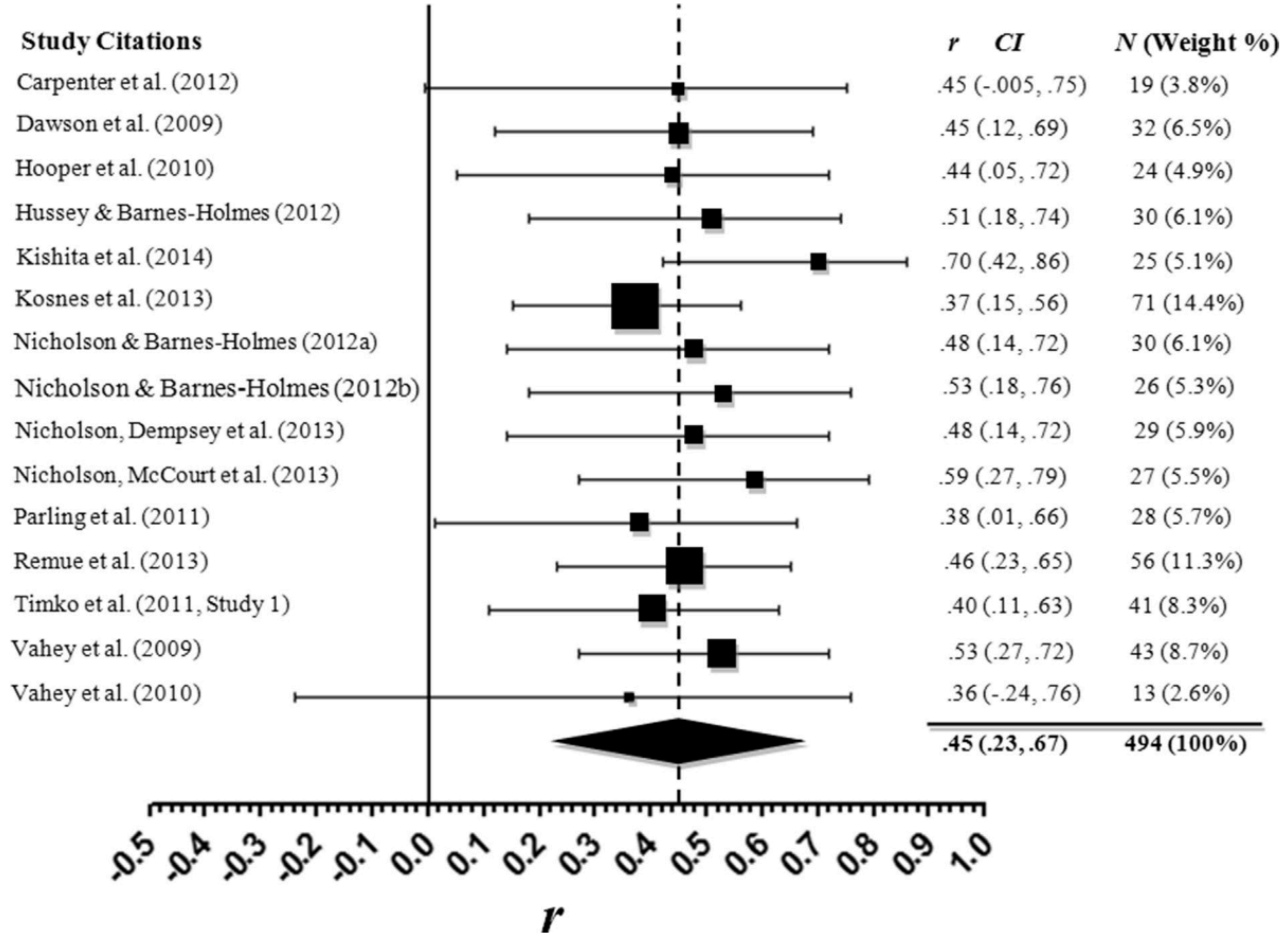
Reported forest plot and funnel plot



“Information that is repeated or overlooked”

Compare data in plots: they differ

Forest plot: weighted-average effect sizes & meta-analysis results



Analytic method: "Hunter and Schmidt meta-analysis"

Code: maybe from Field & Gillett (2010) ??

Forest plot: weighted-average effect sizes & meta-analysis results

				95% CI		95% CR		95% PI	
Source	Implementation	Modifications from the original code	$\bar{r}$	Lower	Upper	Lower	Upper	Lower	Upper
Vahey et al. (2015)	Vahey et al. (2015) state they followed Field & Gillett's (2010) description of a Hunter and Schmidt style meta-analysis	Unknown.	.45	.40	.54	.23	.67	-	-
Verification attempt 1	Hunter & Schmidt method using Field & Gillett's (2010) "h_s_syntax.sps"	All reliabilities were set to 0.	.47	.20	.74	.47	.47	-	-
Verification attempt 2	Hunter & Schmidt method using Field & Gillett's (2010) "Meta_Basic_r.sps" *	Set variance in population correlations to zero if it is negative so that CRs must be non-negative.	-	-	-	-	-	-	-
Verification attempt 3	Hunter & Schmidt method using a reimplementation of Field & Gillett's (2010) "Meta_Basic_r.sps" in R	Set variance in population correlations to zero if it is negative so that CRs must be non-negative.	.46	-	-	.46	.46	-	-
Verification attempt 4	Hunter & Schmidt method using a converstion of Field & Gillett's (2010) "Meta_Basic_r.sps" to R	Set variance in population correlations to zero if it is negative so that CRs must be non-negative. Removed erroneous Overton transformations.	.47	-	-	.47	.47	-	-
Verification attempt 5	Hunter & Schmidt method using Viechtbauer's (2022) implementation in R and metafor.	Credibility intervals were implemented using Field & Gillett's (2010) equations 2 to 5.	.47	.40	.54	.47	.47	.40	.54
Verification attempt 6	A mix of Hunter & Schmidt and Hedges methods using Viechtbauer's (2022) implementation in R and metafor.	Credibility intervals were implemented using Field & Gillett's (2010) equations 2 to 5. Fisher's r-to-z transformations and z-to-r back transformations.	.47	.40	.54	.47	.47	.40	.54

Vahey's data + Vahey's code != Vahey's results

Forest plot: weighted-average effect sizes & meta-analysis results

Source	Implementation	Modifications from the original code	$\bar{r}$	95% CI		95% CR		95% PI	
				Lower	Upper	Lower	Upper	Lower	Upper
Vahey et al. (2015)	Vahey et al. (2015) state they followed Field & Gillett's (2010) description of a Hunter and Schmidt style meta-analysis	Unknown.	.45	.40	.54	.23	.67	-	-
Verification attempt 1	Hunter & Schmidt method using Field & Gillett's (2010) "h_s_syntax.sps"	All reliabilities were set to 0.	.47	.20	.74	.47	.47	-	-
Verification attempt 2	Hunter & Schmidt method using Field & Gillett's (2010) "Meta_Basic_r.sps" *	Set variance in population correlations to zero if it is negative so that CRs must be non-negative.	-	-	-	-	-	-	-
Verification attempt 3	Hunter & Schmidt method using a reimplementation of Field & Gillett's (2010) "Meta_Basic_r.sps" in R	Set variance in population correlations to zero if it is negative so that CRs must be non-negative.	.46	-	-	.46	.46	-	-
Verification attempt 4	Hunter & Schmidt method using a converstion of Field & Gillett's (2010) "Meta_Basic_r.sps" to R	Set variance in population correlations to zero if it is negative so that CRs must be non-negative. Removed erroneous Overton transformations.	.47	-	-	.47	.47	-	-
Verification attempt 5	Hunter & Schmidt method using Viechtbauer's (2022) implementation in R and metafor.	Credibility intervals were implemented using Field & Gillett's (2010) equations 2 to 5.	.47	.40	.54	.47	.47	.40	.54
Verification attempt 6	A mix of Hunter & Schmidt and Hedges methods using Viechtbauer's (2022) implementation in R and metafor.	Credibility intervals were implemented using Field & Gillett's (2010) equations 2 to 5. Fisher's r-to-z transformations and z-to-r back transformations.	.47	.40	.54	.47	.47	.40	.54

Vahey's data + Vahey's code != Vahey's results

Reproduce meta-analysis: not reproducible

Effect size weightings

Individual correlations vs weighted-average correlations

Reported individual correlations:	.82, .58, .44
Reported weighted average correlation:	0.53
Recalculated weighted average correlation:	0.58

Effect size weightings

Individual correlations vs weighted-average correlations

Reported individual correlations:	.82, .58, .44
Reported weighted average correlation:	0.53
Recalculated weighted average correlation:	0.58

Overall: 2 of 15 wrong

Reproduce effect size weighting: not reproducible

Effect size conversions

From Cohen's *d*, η_p^2, etc. to Pearson's *r*

1. Check if reported ones match recalculated ones

2. Check if conversions are possible

η^2 (eta squared) <u>can</u> be converted to Pearson's *r*
η_p^2 (partial eta squared) <u>cannot</u> be converted to Pearson's *r*

Reproduce effect size conversions: some errors

Assess incorrect inclusions

Compare inclusion criteria vs. included effect sizes

“a meta-analysis of clinically-focused IRAP effects …
against their respective criterion variables”

correlation(IRAP, clinical criterion)

Assess incorrect inclusions

Compare inclusion criteria vs. included effect sizes

"a meta-analysis of clinically-focused IRAP effects …
against their respective criterion variables"

Study Citations, IRAP effects and their Involvement in Criterion Relationships	Original Metric with Degrees of Freedom	Resulting r Metric[#]
Carpenter, Martinez, Vadhan, Barnes-Holmes & Nunes (2012)		
'With-cocaine-positive' trial-type D_{IRAP} at baseline		
Negative correlation with number of vouchers earned by cocaine addicts' during first phase of treatment for cocaine addiction.	$r(17) = -.46$	.46
Negative correlation with cocaine addicts' percentage attendance during first phase of treatment for cocaine addiction.	$r(17) = -.58$	.58
Negative correlation with cocaine addicts' percentage of negative urine tests for cocaine use during first phase of treatment for cocaine addiction.	$r(17) = -.56$	.56
'With-cocaine-negative' trial-type D_{IRAP} at baseline		
Negative correlation with number of vouchers earned by cocaine addicts' during first phase of treatment for cocaine addiction.	$r(17) = -.38$	.38
Negative correlation with cocaine addicts' percentage attendance during first phase of treatment for cocaine addiction.	$r(17) = -.48$	.48
Negative correlation with cocaine addicts' percentage of negative urine tests for cocaine use during first phase of treatment for cocaine addiction.	$r(17) = -.48$	.48
'No-cocaine-negative' trial-type D_{IRAP} at baseline		
Negative correlation with number of vouchers earned by cocaine addicts' during first phase of treatment for cocaine addiction.	$r(17) = -.47$	.47
Negative correlation with cocaine addicts' percentage attendance during first phase of treatment for cocaine addiction.	$r(17) = -.23$	.23
Negative correlation with cocaine addicts' percentage of negative urine tests for cocaine use during first phase of treatment for cocaine addiction.	$r(17) = -.40$	.40
Dawson, Barnes-Holmes, Gresswell, Hart & Gore (2009)		
'Child-sexual' trial-type D_{IRAP}		
Positive effect between controls versus paedophilic sex-offenders.	$F(1, 30) = 6.136$	.41[a]
Positive one-group effect among controls.	$d(15) = 1.34$*	.56[b]

Assess incorrect inclusions

Compare inclusion criteria vs. included effect sizes

"a meta-analysis of clinically-focused IRAP effects …
against their respective criterion variables"

Study Citations, IRAP effects and their Involvement in Criterion Relationships	**Original Metric with Degrees of Freedom**	**Resulting *r* Metric[#]**
Carpenter, Martinez, Vadhan, Barnes-Holmes & Nunes (2012)		
'With-cocaine-positive' trial-type D_{IRAP} at baseline		
Negative correlation with number of vouchers earned by cocaine addicts' during first phase of treatment for cocaine addiction.	$r(17) = -.46$	.46
Negative correlation with cocaine addicts' percentage attendance during first phase of treatment for cocaine addiction.	$r(17) = -.58$	.58
Negative correlation with cocaine addicts' percentage of negative urine tests for cocaine use during first phase of treatment for cocaine addiction.	$r(17) = -.56$	.56
'With-cocaine-negative' trial-type D_{IRAP} at baseline		
Negative correlation with number of vouchers earned by cocaine addicts' during first phase of treatment for cocaine addiction.	$r(17) = -.38$	.38
Negative correlation with cocaine addicts' percentage attendance during first phase of treatment for cocaine addiction.	$r(17) = -.48$	.48
Negative correlation with cocaine addicts' percentage of negative urine tests for cocaine use during first phase of treatment for cocaine addiction.	$r(17) = -.48$	.48
'No-cocaine-negative' trial-type D_{IRAP} at baseline		
Negative correlation with number of vouchers earned by cocaine addicts' during first phase of treatment for cocaine addiction.	$r(17) = -.47$	.47
Negative correlation with cocaine addicts' percentage attendance during first phase of treatment for cocaine addiction.	$r(17) = -.23$	.23
Negative correlation with cocaine addicts' percentage of negative urine tests for cocaine use during first phase of treatment for cocaine addiction.	$r(17) = -.40$	.40
Dawson, Barnes-Holmes, Gresswell, Hart & Gore (2009)		
'Child-sexual' trial-type D_{IRAP}		
Positive effect between controls versus paedophilic sex offenders.	$F(1, 30) = 6.136$	.41[a]
Positive one-group effect among controls.	$d(15) = 1.34$*	.56[b]

Assess incorrect inclusions

Compare inclusion criteria vs. included effect sizes

“a meta-analysis of clinically-focused IRAP effects …
against their respective criterion variables”

correlation(IRAP, clinical criterion)

Involves criterion variable	*N* effect sizes
TRUE	33
FALSE	23

Check correct inclusions: many errors

Assess incorrect <u>exclusions</u>

Compare inclusion criteria vs. excluded effect sizes

Effect sizes meeting criteria	*N*
Reported	56

Assess incorrect <u>exclusions</u>

Compare inclusion criteria vs. excluded effect sizes

Effect sizes meeting criteria	*N*
Reported	56
Reextracted	171

Check correct exclusions: many errors

Assess incorrect exclusions

Compare inclusion criteria vs. excluded effect sizes

TABLE 2. Pearson's correlation coefficients testing the relationship between performance (implicit)-based measures.

	Treatment Weeks 1–12		
Measures	Voucher earnings	Visits attended (%)	UA cocaine negative (%)
D-IRAP scores by trial type[3]			
With Cocaine Positive	−.46*	−.58**	−.56*
With Cocaine Negative	−.38	−.48*	−.48*
No Cocaine Positive	−.03	−.19	−.18
No Cocaine Negative	−.47*	−.23	−.40

Assess incorrect <u>exclusions</u>

Compare inclusion criteria vs. excluded effect sizes

TABLE 2. Pearson's correlation coefficients testing the relationship between performance (implicit)-based measures.

Measures	Treatment Weeks 1–12		
	Voucher earnings	Visits attended (%)	UA cocaine negative (%)
D-IRAP scores by trial type[3]			
With Cocaine Positive	−.46*	−.58**	−.56*
With Cocaine Negative	−.38	−.48*	−.48*
No Cocaine Positive	−.03	−.19	−.18
No Cocaine Negative	−.47*	−.23	−.40

Mean r = .45

Mean r = .13

Check correct exclusions: many errors / selection bias

Assess incorrect exclusions

Compare inclusion criteria vs. excluded effect sizes

TABLE 2. Pearson's correlation coefficients testing the relationship between performance (implicit)-based measures.

	Treatment Weeks 1–12		
Measures	Voucher earnings	Visits attended (%)	UA cocaine negative (%)
D-IRAP scores by trial type[3]			
With Cocaine Positive	−.46*	−.58**	−.56*
With Cocaine Negative	−.38	−.48*	−.48*
No Cocaine Positive	−.03	−.19	−.18
No Cocaine Negative	−.47*	−.23	−.40

Mean r = .45

Mean r = .13

Check correct exclusions: many errors / selection bias

Compound impact of errors

Reported: $\bar{r}$ = .45, 95% CI [.40, .54], 95% CR [.23, .67]

New: $\bar{r}$ = .22, 95% CI [.15, .29], 95% PI [-.01, .42]

			Vahey et al. (2015)		Verified	New meta-analysis	
Test	Tails	Estimated using*	$\bar{r}$	N	N	$\bar{r}$	N
Pearson's r	One	Point estimate	0.45	29	29	.22	126
Pearson's r	One	Lower bound of 95% CI	0.40	37	37	.15	273
Pearson's r	Two	Point estimate	0.45	36	36	.22	160
Pearson's r	Two	Lower bound of 95% CI	0.40	-	46	.15	346
Independent t-test (Cohen's d)**	One	Point estimate	1.01	26	26	.45	124
Independent t-test (Cohen's d)**	One	Lower bound of 95% CI	0.87	36	34***	.30	270
Dependent t-test (Cohen's d) **	One	Point estimate	1.01	8	8	.45	32
Dependent t-test (Cohen's d) **	One	Lower bound of 95% CI	0.87	10	10	.30	69

No published studies have reached these Ns!

Questions?