Do psychological effects fade away?

Meta-analyzing the decline effect.

Lisa Wijsen

University of Amsterdam

January 15, 2020

Introduction

Over the last years, we have learned a great deal about the reliability of psychology and other scientific disciplines. Major fraud cases, like Diederik Stapel's in 2011, have brought to light that (psychological) science may not always be as reliable as we thought before and that replication studies are a necessity (?, ?). Besides fraud cases, other factors have given rise to questions on the reliability of science as well. For example, the rate of false positives in scientific literature seems to be too high (?, ?, ?), suggesting that many scientific findings are simply not true (?, ?, ?).

It is also argued that psychological research articles incorporate an excess of significant results (?, ?, ?, ?). One explanation for the excess of significant results is publication bias: scientific journals favour studies that find significant results, which are then more likely to get published. As publication bias impedes certain studies from being published (e.g. studies that find no effect), it can lead to a biased view of all research that is actually being done. Moreover, replications of studies do not always corroborate earlier findings, like those presented by ? (?) on looking into the future (?, ?). All in all we can argue that published scientific findings do not always have a reliable claim on the truth, that effects can be inflated and perhaps even that 'most published results are false' (?, ?). The outcome of one study is not always a reliable representation of its effect in 'the real world' and further research is always necessary in order to draw right conclusions.

The unreliability of scientific findings is strengthened by what is supposed to be the 'decline effect' (?, ?) which states that effects that have once been found sometimes become smaller or fade away over the years. If the decline effect is true, one expects earlier studies to find larger effect sizes than later studies. This may not only count for the false positives - finding effects that are indeed non-existent or false -, but also for true effects (?, ?). Besides all the factors already mentioned, the decline effect would be another reason that explains why scientific findings are not always reliable and why extra attention should be paid to replication studies: effect sizes from early studies could be overestimated. This study will be the first look into the decline effect in psychology.

Researchers have formulated several explanations for the existence of the decline effect. The most popular explanation for the decline effect is publication bias (?, ?, ?). Once researchers find write an article that involves one or more significant effects, the study is often published. The publication of studies that find smaller effects, or do not find the effect at all, is delayed or simply does not take place (?, ?, ?). Moreover, it is easier to publish confirmatory results in an early stage of research in a specific field than later on, when more critical theories have been developed (?, ?). Publication bias could explain the decline effect, as initial positive effects are published sooner than studies that refute this outcome, which means that there is a period of time in which the average effect in a meta-analysis of available studies stabilizes to a lower and possibly more accurate value. Evidence for the existence of publication bias has been found in different scientific fields, among which the medical sciences (?, ?) and psychological science (?, ?) and could therefore be a viable explanation for the fading of effects.

A second possible explanation - the increased sample explanation - entails that the first study within a new scientific field is smaller in terms of sample size than studies that are done and published later on (?, ?). Studies with a small sample size are underpowered and are therefore less likely to find an effect given that it exists in reality. When a small study results in a significant p-value, one cannot be sure whether it respresents a true effect or whether it indicates a false positive. Once these smaller studies are completed and published, researchers collect larger samples and might also incorporate other moderators in their study. In this case, it is more likely that these bigger studies find smaller effect sizes than the initial, smaller studies, because the underlying effect is indeed smaller. This does not necessarily have anything to do with publication bias: it is simply how science works. Once an effect has been found in a small study, other, larger, studies will be run to replicate this finding, but these replications will also be aimed to relate the original effect to particular moderators or other variables that are related to the specific effect. Adding these moderators might also lead to a decline in effect sizes over the years, as variance that was first explained by one variable is now explained by multiple variables. Larger sample sizes and the addition of extra moderators could all together cause the decline effect. What distinguishes this scenario from one in which publication bias is the malefactor, is that the underlying true effect is indeed smaller and effect sizes are not biased due to the selection mechanism for publishing studies.

A third possible explanation for the decline effect is that effects do indeed fade away over the years (the true decline effect explanation). Usually this is not a valid assumption, as an effect that is true will stay true over time (?, ?), yet there are possible scenarios in which this is not always be the case. For instance, one may argue that a psychological effect that has once been found fades away over time, because participants of psychological experiments have become conscious of this effect. For example: because participants in a study are often psychology students, they could be aware of a certain effect that the researchers intend to measure. Researchers are no longer able to measure the full effect and as awareness usually increases over time, later studies will find smaller effect sizes.

Evidence for the existence of the decline effect has been found in the medical sciences (?, ?) and in the biological sciences (?, ?). ? (?) assessed the decline effect by performing a meta-analysis on the effect of four different medications and found fading effect sizes in three out of four types of medications. ? (?) performed a so-called meta-meta-analysis in which they analyzed different meta-analyses in the domain of biology and ecology and also found a small but significant negative relation between year of publication and effect size, meaning that earlier studies indeed found larger effect sizes than studies that were executed later on. In accordance with ? (?), we will also use a meta-meta-analytical approach in our study, meaning that we will analyze multiple meta-analyses to investigate the existence of the decline effect in psychological science. A meta-meta-analysis is a powerful statistical technique that enables us to draw general conclusions on trends in effect sizes over a substantial number of primary studies (more on the meta-meta-analysis can be found in the methods section).

Both the medical sciences and the biological sciences focus less on the measurement of behavioural variables than psychology does, whereas it is exactly the focus on behavioural measurement that can lead to extreme effect sizes (?, ?). Verifying the existence of the decline effect in psychological science would therefore be of utmost importance. If the decline effect indeed is common within psychology, conclusions drawn from early results have to be adjusted or nuanced at the least. Knowledge on the decline effect is therefore essential if we want to interpret psychological research correctly.

Within psychology, no research has been done on the existence of decline effect, therefore this study is mainly of an exploratory nature. In this study, we will investigate the existence of the decline effect in psychological science and, if it exists, aim to find evidence for either the publication bias explanation, the increased sample size explanation or the true decline effect explanation. We will do this through, as mentioned before, a meta-meta-analysis. Through a meta-meta-analysis, we can investigate whether effect sizes fade away over time, not only in one specific domain, but over several domains. We will be able to systematically review temporal trends in effect sizes as a sample of meta-analyses will cover a large total of primary studies, and thereby give a reliable overall view on the decline effect and possible explanations.

Methods

Selection of Meta-Analyses and Data Sources

We drew a random sample of 23 meta-analyses of all meta-analyses published in 2011 and 2012 that were made available through search engine PsycInfo, and specifically in PsycArticles. As a search criterium, we used 'meta-a*' which resulted in 129 hits for 2011 and 118 for 2012. For a meta-analysis to be suitable for this study, it had to entail a table with each primary study, and its corresponding year of publication, effect size and its standard error or within-variance. We only sampled meta-analyses that made use of either a correlation effect size or a standardised mean effect size (Cohen's d or Hedges' g). Moreover, each meta-analysis had to cover at least 10 primary studies.

Of the 247 references, 98 were not a proper meta-analysis but were for example of a methodological nature or incorporated a term similar to meta-analysis. There was a total of 97 meta-analyses that we could not use due to lack of data in the meta-analytic papers (either the effect size, its standard error and/or year of publication of each primary study was missing). Fifty-two meta-analyses were indeed usable, of which we drew a random sample of 23 meta-analyses. To our sample we added another 11 meta-analyses that were used in previous research (?, ?), resulting in a total of 34 meta-analyses. The meta-analyses involved a total of 2729 primary studies. These meta-analyses came from a variety of journals within the field of psychological science (e.g.*Journal of Applied Psychology*, *Psychological Bulletin*, *Psychotherapy*, *Journal of Counseling Psychology*, *Journal of Educational Psychology*, *Journal of Personality and Social Psychology*). With this sample, we believe that we cover a representative part of psychology insofar that the studies were incorporated in meta-analyses on psychological topics, and by investigating this data systematically, we will in the end be able to draw reliable conclusions on the existence of the decline effect within psychological science.

Data Extraction

From each primary study in the sample, the year of publication, the effect size and standard error were extracted or computed. When a meta-analysis involved at least one categorical moderator, we constructed multiple subsets for each of these meta-analysis to avoid heterogeneity within each subset. We set a minimum of 10 studies for each subsets to ensure that the power was high enough to estimate the average effect sizes accurately. If there was more than one categorical moderator included in the meta-analysis, we chose the first one mentioned in the results. So for instance, ? (?) meta-analyzed the prediction of behavioral, judgment and physiological measures by the Implicit Association Test. A categorical variable that ? (?) incorporated in the study was 'topic', which means that the authors drew a distinction between different topics in terms of predicting behavioral, judgment and physiological measures. We divided our data in subsets, each only involving those studies that measured a variable on that specific topic. We did not use ? (?) in our analysis, as it does not come from 2011 or 2012, but we used a similar method to construct subsets for each of the meta-analyses in our sample(if such categorical moderators were present in the data). To overcome dependency in the data, we made sure that each primary study was only used once in the entire meta-meta-analysis: each subset was completely independent of others, and when a primary study was used more than once in a meta-analysis, we selected only the first one mentioned in the table in the meta-analysis. This was only the case when it was clear that the same sample of people was used multiple times in the meta-analysis; when one study drew different samples for different experiments, we included each one of these in the analysis.

Statistical Analysis

We used the R statistical software 3.0.2 (R Development Core Team, 2013), and specifically the package 'metafor' for the meta-analytical analyses (?, ?). As the type of effect size varied among studies, we converted each to Hedges' g (a standardized mean effect size measure that controls for the slight positive bias of the Cohen's d effect size (?, ?)). We also recoded effect sizes whenever a negative effect size indicated a larger effect, to make sure that all effect sizes were coded in the right direction (a higher effect size means a larger effect).

We fitted several random effect models. The choice for a random effect model instead of a fixed effects model comes from the fact that we find it more likely that the true effect size is not the same for each and every study that investigates the same variables: it is likely that variations among effect sizes are possible that are not due to error. To give a simple example: the effect of a specific therapy might be stronger for children than for students, resulting in varying effect sizes for these two different groups of participants.

We firstly fitted a random effects model on each subset to estimate the average effect size. In a random effects model, the effect size is regressed on the grand mean (μ), the deviation from the true effect (ζ_i) as we allow variation in the true effect size, and an error term (ϵ_i). Both fixed effects models as random effects models weight the effect size according to sample size, but as the random effects models allow more variation in the underlying true effect size (they treat it as being random), the influence of the sample size is weaker for random effects models than for fixed effects models.

$$Y_i = \mu + \zeta_i + \epsilon_i \tag{1}$$

To investigate the existence of the decline effect, we fitted a random effect model on each subset with year of publication (*yop*) as a moderator - the effect size is regressed on year of publication-, to investigate whether year of publication leads to lower effect sizes.

$$Y_i = \beta_0 + \beta_1 yop_i + \epsilon_i \tag{2}$$

Thirdly, we repeated this procedure but with the standard error as a predictor (*se*). Using the standard error as a predictor for the effect size, is one way to test for funnel plot asymmetry, and is often called the Egger test (?, ?, ?). A funnel plot shows effect size estimates on the horizontal axis and sample size (or the standard error) on the vertical axis. A funnel plot is symmetrical when the larger studies are roughly in the middle and the smaller studies scattered around it. In presence of publication bias, this plot is often found to be asymmetrical, as studies finding smaller effect sizes are missing. To illustrate the difference between an asymmetrical and a symmetrical funnel plot, see Figure 2. These are funnel plots from two subsets from ? (?). The first subset, 'consumer', is not suffering from publication bias as all studies (all the dots) are evenly scattered in the plot. The clinical subset on the other hand, does show a lack of small studies with low effect sizes, which indicates the possibility of publication bias.

When the standard error, a measure of precision or sample size, is indeed a predictor, it means that the funnel plot is not symmetrical like in the clinical example, which can then in turn be a sign of publication bias. A small cautionary note: although publication bias is indeed a possible explanation for the found lack of small studies with a low effect size (or in other words, the asymmetry of funnel plots), other causes for heterogeneity are also possible (?, ?). For example, studies that investigate a specific type of participants (e.g. people suffering from a rare disorder or disease) are often small and find high effect sizes as these types of participants are more susceptible to medication or therapy than the average participant. It was beyond the scope of this study to verify whether asymmetry of funnel plots was indeed due to publication bias or some other cause, but for now we will assume that publication bias is the most likely explanation, as its existence in psychology has been shown before (?, ?, ?).

$$Y_i = \beta_0 + \beta_1 se_i + \epsilon_i \tag{3}$$

Fourthly, we ran another random effects model with both year of publication and the standard error as predictors. With this model, we can find out whether the decline effect also persists even when we control for publication bias. As mentioned earlier, a possible explanation for the decline effect is publication bias. In this model we test whether the decline effect persists, even when we control for publication bias. If year of publication is no longer a predictor for the effect size when the standard error is included in the model, we can conclude that the decline efect can be explained by publication bias. If year of publication is still a significant predictor eventhough the standard error is included in the model, the conclusion can be drawn that publication is not an explanation for the decline effect, and that there are possible other moderators at hand that explain declining effect sizes.

$$Y_i = \beta_0 + \beta_1 yop_i + \beta_2 prec_i + \epsilon_i \tag{4}$$

Up till now we ran several random effects models on each subset separately, and calculated the measures and coefficients mentioned above, but first and foremost we are interested in the overall weighted measures across all subsets. We therefore performed several meta-meta-analyses: we ran several random effects models on i) the average effect size of each subset, ii) the slope of year of publication on effect size, iii) the slope of the standard error on effect size, iv) the slope of year of publication on effect size when controlling for the standard error and v) the slope the standard error on effect size when controlling for year of publication and vi) the correlation between year of publication and sample size. Through applying a powerful technique such as meta-meta-analysis, we obtained overall weighted estimates of all the measures mentioned across all subsets.

Results

In total, we analyzed 58 subsets of primary studies, extracted from 34 meta-analyses. These covered a total of 2729 primary studies, all within the field of psychological science. Table 1 shows for each of these subsets the average effect size and its standard error, the effects of year of publication and of the standard error, the correlation between year of publication and the standard error, and lastly the number of primary studies per meta-analysis.

The average sample size N per meta-analysis was 47, ranging from 10 (the minimum we required for studies to be included in the analyses) to 297. The overal average effect size in Hedges'g was estimated at 0.3706 with a standard error of 0.0728 (CI[0.2279, 0.5133]), with a minimum of -0.1437 and a maximum of 1.776. Figure 2 shows the distribution of effect sizes among the 58 subsets. As 0 does not lie between the lower and upper bounds of the confidence interval, we can conclude that there is indeed an overall significant positive effect size, which could be interpreted as a moderate or medium effect. This is consistent with what has been found earlier in psychology (?, ?). Forty-three out of 58 subsets tested significantly on the Q-test for heterogeneity, indicating that in these subsets the measured effect size varied significantly over the primary studies (see Table 1).

The average effect of year of publication on the effect size was -0.0004, with a standard error of 0.0008 (CI[-0.0020, 0.0012]) with a minimum of -0.0717 and a maximum of 0.1082. Although the average effect has indeed a negative direction, which implies that later studies produce smaller effect sizes, the confidence interval contains 0 and we can therefore conclude that the overall decline effect is not significant. Out of all subsets, year of publication was a significant predictor for effect size in three subsets (?, ?, ?, ?). In two of these subsets, year of publication was no longer significant predictor for the effect size when controlling for the standard error. This would indicate that the decline effect in these two subsets dissapears when controlling for publication bias, meaning that the decline effect in these sets is possibly caused by the lack of published studies with low effect sizes. Interestingly, the opposite effect (effect sizes growing larger over the years) was also found in three of the susbsets (see table 1). Figure 3 shows the distribution of the decline effect and its confidence interval. These results indicate that we have no overall evidence for the existence of the decline effect, but that it might occur in specific fields.

To assess publication bias in all subsets, we added the standard error as a moderator for the effect size in the meta-meta-analysis. An average positive effect of the standard error would imply that there is a lack of small studies that find small effect sizes or no effect at all (the funnel plot of such a subset would be skewed). The average 'publication bias' effect across the 58 subsets was 0.7527 with a standard error of 0.1369 (CI[0.4844, 1.0220]), with a minimum of -1.9226 and a maximum of 5.9321. As the confidence interval (see Figure 4) did not contain the 0, we can conclude that there is a positive overall effect of the standard error on the effect size, indicating an overal indication of publication bias. Out of the 58 subsets, we found a positive effect of the standard error on the effect size in 11 cases (see table 1). In these subsets, the effect size increases when the standard error increases (or the sample size decreases) which indicates a skewed funnel plot: these subsets lack studies with low effect sizes.

We performed two meta-analyses to investigate the effect of year of publication and of the standard error of the effect size, when combined in the same regression. Year of publication had an effect on the effect size of 0.0000, with a standard error of 0.0008, and the standard error had an average effect of 0.5844 and a standard error of 0.1309. These values do not deviate much from those calculated earlier: again the decline effect is now practically zero, and the effect of the standard error is significant and positive. The effect of year of publication no longer showed a negative direction, which could be due to the fact that we controlled for the standard error. Across the board, it appears to be that the slight decline effect can be explained by publication bias as evidenced by asymmetry in the funnel plots. Nevertheless, the decline effect was also non-significant without controling for the standard error, indicating that the decline effect is not substantial to begin with.

Lastly, we calculated the correlation between year of publication and the standard error, to investigate whether sample size increases over the years. The mean correlation between year of publication and the standard error was -0.1077, with a standard error 0.0393. A negative correlation implies that the later a study is published, the smaller its standard error is, and the larger its sample size. The correlation is significant, but represents a small effect. Figure 5 shows the distribution of the found correlations in the 58 subsets. We can conclude that eventhough sample sizes increased over the years, effect sizes do generally not decline as time passes by.

Conclusion and Discussion

In this study, we have systematically investigated the existence of the decline effect in psychological science, using 58 subsets from 34 meta-analyses, covering a total of 2729 studies. Using meta-analytical techniques we aimed to establish the overall presence of the decline effect, publication bias and the increase of sample sizes throughout the years.

Although its direction was slightly negative, we can conclude that there is no overall decline effect in the studied meta-analyses. Overall, effect sizes do not appear to decrease over the years. The effect size averaged over 58 subsets was 0.37, indicating that generally effect sizes in psychology have a medium size, thereby corroborating earlier summaries (?, ?, ?, ?, ?, ?, ?). As ? (?) have shown, assuming a median effect size combined with a sample size of 40 (close to the average of 47 participants we have found), results in low power. This indicates that generally, psychological studies have too small sample sizes (and thus too low power) to detect psychological effects given that they exist. As there is an abundance of positive results in the literature (?, ?, ?, ?), it is highly possible that publication bias is indeed a problem in psychological science.

In this study, we found that the standard error is indeed a predictor for effect size, which indicates that small studies with small or null-effect sizes are generally missing (possibly indicating publication bias). When controlling for publication bias, we saw that the negative direction of the decline effect turned neutral. We also found an overall negative correlation between year of publication and the standard error indicating that sample sizes increase over time. In conclusion, eventhough we mentioned both publication bias and increased sample sizes as possible explanations for the decline effect, we have found no overall decline effect in the field of psychological science.

The lack of any clear decline effect is of course good news: apparently old studies are in terms of effect size estimation as reliable as current studies and there is no need for correcting these estimates. But eventhough we have found no clear overall decline effect, this study does not deny the possibility of the existence of the decline effect within specific fields. It is still possible that the decline effect is present in for instance the parapsychology as ? (?) argued or some other field. Nevertheless, our results show that the decline effect is not a general trend within psychology and so suggest that it is is not a major issue. Our results do confirm that publication bias, or the lack of small studies with small effect sizes, is indeed a trend within psychology, and one that is worrisome. Due to the lack of published studies with small effect sizes, meta-analyses can overestimate average effect sizes if they do not control for publication bias in their analysis. Therefore our estimated average effect size of 0.37 might be a slight overestimation as well. Again, we must be cautious here: the standard error being a predictor for the effect size (or asymmetrical funnel plots) can be a sign of publication bias, but might also have other causes (?, ?). It was beyond the scope of this article to investigate whether our found effect might be due to publication bias or not, but as publication bias is prominently present in psychology, we for now assume that publication bias is indeed an adequate explanation. Whichever explanation holds: it is important to stress that in terms of effect size estimates, we rather put our faith in the larger studies, as these tend to be more precise than in small studies that are either not published or overestimate effect sizes.

This study and mainly the data that we have collected have given us many possibilities for further research. A topic that was beyond the scope of this article but certainly of interest is for instance statistical power analysis. Psychological studies appear to be structurally underpowered (?, ?) due to for instance small sample sizes. The data we have collected enables us to perform these power analyses, and systematically verify whether psychological studies are indeed underpowered. Other topics that are interesting for further research are for instance the differences between American and non-American publications in terms of effect size, as American studies seem to find higher effect sizes than studies of a non-American origin (?, ?). Obviously, other geographical differences can also be taken into account. A third possible topic of research is the contrast between different fields within and outside psychology. In this study we have made no distinction between different subdisciplines (social psychology, clinical psychology etc.), whereas certain effects within specific disciplines might be more vulnerable to the decline effect than others. For instance, the effect of a certain type of psychotherapy might possibly remain constant over the years, but a social psychological effect, such as priming, that is predominantly studied among pschology students might fade away over time. This study has only provided a general conclusion to a very general question, but variation is possible between different disciplines or geographical areas.

One question that remains is why there would not be any decline effect if there is indeed publication bias and an increase of sample sizes. This is a hard question to answer, as the assumption that both lead to the fading of effect sizes does not entirely emerge from thin air. It is reasonable to believe that both indeed lead to the decline effect, and now that they do not behave according our hypotheses, we need an alternative explanation. An explanation for the increased sample sizes is that the overall correlation we found between year of publication and standard error was significant, but low. Sample sizes increase over the years, but may not increase enough to cause effect sizes to fade away over time. Perhaps when sample sizes will increase more strongly, the decline effect would pop up. For publication bias, we might have to think in a different direction. In this study we have assumed that publication bias causes the decline effect by delaying the publication of studies with low effect sizes. As this takes a while, later studies might have lower effect sizes than earlier studies. But we can argue in another direction as well: what if the publication bias is still withholding studies with small effect sizes from being published? In this case, the studies that would in the end cause the decline effect are still not available to the larger public, and we simply cannot estimate correctly whether there is a decline effect or not. It would be interesting to investigate whether the decline effect varies in strength between meta-analyses that study relatively 'old' problems and studies that incorporate relatively recent questions.

Two more topics that need to be addressed are the representativity of our sample and the heterogeneity of effect sizes within our subsets. Whether our sample of meta-analyses is indeed representative will always remain a matter of subjective estimation. The field of psychological science has become rather substantial and obviously, no sample could ever cover its entirety. Moreover, it is possible that certain areas are more keen on performing meta-analyses than others and are therefore over- or underrepresented. Nevertheless, our sample nearly covers a total of 3000 primary studies covering different fields within psychology (e.g. clinical psychology, neuropsychology, educational psychology, social psychology), and we may cautiously conclude that our sample provides us with a very general idea of the decline effect in psychological science.

In this study, we tried to ensure homogeneity in each subset by constructing subsets that controlled for possible moderators. Through these method, we hoped to aggregate subsets which variance was caused by the main effect, and not by other moderators that were of interest. A measure for heterogeneity in a sample is the Q-test (?, ?), which quantifies the heterogeneity of the underlying effect size. As we can see in Table 1, most subsets test significantly for heterogeneity of effect sizes which suggests that the underlying effect size varies within each meta-analysis. As we have chosen for random effects models, we allowed effect sizes to vary within each subset and this might also be an explanation for the found heterogeneity. Nevertheless, we can conclude that the majority of our subsets are not homogeneous which could be due to other moderators that we have not taken into account. In future studies, it must be taken into consideration whether it is indeed possible to include more moderators when constructing subsets in order to construct more homogeneous subsets, and whether it is fruitful to use fixed effects models instead of random effects models, as these assume that the effect size is the same within each meta-analysis.

Despite these limitations, this study provides a systematic review of the existence of the decline effect, which is a first in the history of psychological science. Although it does not deny its existence in specific fields, it indicates that overall psychology does not have to worry about fading effect sizes: they remain stable over time. Over the last years much has been written on the questionable research practices and the lack of good interpretation in psychology, but we can conclude that this decline effect does not distort our findings. We have also seen that publication bias is indeed still a problem, and finding ways to publish every result instead of only those articles that find significant results is of utmost importance, and should be a high priority in psychology. But gladly we can conclude that avoiding the decline effect isn't one.

Table 1: Effect size, effect of Year of Publication (YoP), effect of SE and Correlation (r) between Year of Publication and SE

Meta-Analysis	Q-value (df)	Hedges g (se)	Effect YoP (se)	Effect SE (se)	r YoP, SE	N
Adesope et al. 2012	313.0 (56)*	0.1436 (0.0856)	-0.0030 (0.0121)	0.0608 (0.7970)	-0.5766 (0.0772)	57
Babbage et al. 2012	30.5 (12)*	1.1122 (0.1423)	-0.0035 (0.0228)	0.2243 (0.7475)	0.0410 (0.2871)	13
Balliet et al. 2011, 'mixed'	330.1 (138)*	0.1352 (0.0242)	-0.0002 (0.0002)	0.3568 (0.4016)	-0.0756 (0.0834)	139
Balliet et al. 2011, 'Same'	188.0 (84)*	-0.0853 (0.0362)	0.0065 (0.0024)*	-1.4574 (0.4758)	-0.2079 (0.0996)	85
Berry et al. 2011, 'Hispanic'	205.8 (96)*	0.6287 (0.0239)	0.0020(0.0020)	0.3135 (0.2534)	0.0704 (0.1027)	97
Berry et al. 2011, 'White'	4204.0 (296)*	0.7020 (0.0136)	0.0000 (0.0011)	0.2967 (0.2253)	-0.5398 (0.0388)	297
Berry et al. 2011, 'Asian'	365.3 (58)*	0.6869 (0.0340)	-0.0047 (0.0038)	0.1601 (0.3610)	0.2848 (0.1295)	59
Berry et al. 2011, 'Black'	1411.7 (282)*	0.5444 (0.0185)	0.0046 (0.0015)	0.5566 (0.1828)*	-0.0578 (0.0589)	283
Card et al. 2011	96.7 (15)*	0.1560 (0.1142)	0.0172 (0.0133)	-0.9911 (1.4822)	-0.1812 (0.2244)	16
Crook et al. 2011	262.6 (66)*	0.3932 (0.0396)	0.0001 (0.0003)	0.4080 (0.5345)	0.1406 (0.1243)	67
De Wit et al. 2012	410.4 (94)*	0.1265 (0.0544)	-0.0067 (0.0124)	1.5040 (0.5666)*	-0.0656 (0.1010)	95
Elsequest et al. 2011, 'Non-White'	143.1 (30)*	0.1299 (0.0509)	-0.0011 (0.0049)	-0.3753 (0.8330)	-0.1702 (0.1656)	31
Elsequest et al. 2011, 'Unspecified or mixed'	1144.8 (144)*	0.3091 (0.0387)	-0.0013 (0.0041)	-0.1084 (0.5300)	-0.2270 (0.0761)	145
Elsequest et al. 2011, 'White'	769.2 (130)*	0.2624 (0.0294)	-0.0009 (0.0051)	0.8775 (0.4074)*	0.0578 (0.0881)	131
Fisher et al. 2011, 'Emerg'	250.7 (51)*	0.4873 (0.0947)	0.0025 (0.0092)	2.5904 (0.9622)*	-0.1973 (0.1270)	52
Fisher et al. 2011, 'Non emerg'	143.3 (16)*	0.3707 (0.1467)	-0.0167 (0.0115)	2.0923 (1.6601)	-0.4007 (0.1726)	17
Fox et al. 2011, 'Explanatory'	57.5 (20)*	0.3976 (0.1324)	-0.0317 (0.0098)*	0.3211 (1.5010)	-0.3218 (0.1734)	21

Continued on next page

Table 1 – *Continued from previous page*

Meta-Analysis	Q-value (df)	Hedges g (se)	Effect YoP (se)	Effect SE (se)	r YoP, SE	N
Fox et al. 2011, 'Think Aloud'	56.2 (41)	-0.0763 (0.0575)	0.0002 (0.0004)	0.4899 (0.6524)	0.0780 (0.1577)	42
Fox et al. 2011, 'Directed'	29.7 (10)*	0.1432 (0.1803)	0.0045 (0.0135)	-0.6793 (1.6998)	-0.2914 (0.2384)	11
Freund et al. 2012	242.6 (54)*	0.7577 (0.0585)	-0.0063 (0.0034)	1.7362 (0.6582)*	-0.6191 (0.0717)	55
Ihle et al. 2012	15.4 (19)	0.0304 (0.0214)	-0.0050 (0.0100)	0.2595 (0.3220)	-0.2108 (0.1973)	20
Koenig et al. 2011	1896.6 (76)*	1.0391 (0.1117)	0.0107 (0.0114)	5.9321 (1.2078)*	0.1825 (0.1152)	77
Kolden et al. 2011	55.2 (15)*	0.5142 (0.1426)	0.0056 (0.0096)	0.5146 (1.7329)	-0.0485 (0.2464)	16
Langrehr et al. 2011	29.9 (11)*	-0.1437 (0.1115)	0.0152 (0.0425)	-1.0576 (1.8448)	-0.5489 (0.1556)	12
Mol et al. 2011, 'Preschool Chilren'	15.7 (11)	0.6696 (0.0788)	-0.0131 (0.0203)	0.6696 (0.0788)*	0.2217 (0.3190)	12
Mol et al. 2011, 'Grades 1 - 12'	26.0 (17)	0.9925 (0.0677)	-0.0135 (0.0109)	0.1883 (0.8828)	0.3300 (0.2509)	18
Mol et al. 2011, 'Students'	47.7 (17)*	1.3769 (0.1102)	0.0117 (0.0197)	0.3368 (0.9215)	0.2281 (0.2527)	18
Morgan et al. 2012	25.3 (14)*	0.8871 (0.1256)	-0.0429 (0.0170)	1.0386 (0.8162)	-0.2147 (0.2246)	15
Munder et al. 2012	53.5 (23)*	-0.0651 (0.0886)	-0.0200 (0.0230)	-1.4321 (1.0781)	-0.4014 (0.1485)	24
Piet et al. 2012	17.2 (18)	0.4235 (0.0579)	-0.0207 (0.0156)	1.1588 (0.5501)*	0.1228 (0.2417)	19
Smith et al. 2012	656.6 (183)*	0.3569 (0.0210)	-0.0059 (0.0046)	0.2829 (0.3715)	-0.2257 (0.0678)	184
Tilman, 2011	101.8 (40)*	0.8779 (0.0447)	0.0097 (0.0089)	1.1222 (0.6736)	-0.0330 (0.1552)	41
Toosi et al. 2012, 'SS'	114.0 (23)*	0.1239 (0.1018)	-0.0130 (0.0094)	-1.6242 (2.3161)	0.0306 (0.2082)	24
Toosi et al. 2012, 'MS'	27.9 (9)*	0.2495 (0.1149)	-0.0057 (0.0180)	-1.3944 (1.1447)	0.6713 (0.3057)	10
Van Iddekinge et al. 2011, 'Civilian'	103.5 (39)*	0.3317 (0.0529)	-0.0001 (0.0029)	0.5709 (0.6500)	-0.1294 (0.1502)	40

Continued on next page

Table 1 – *Continued from previous page*

Meta-Analysis	Q-value (df)	Hedges g (se)	Effect YoP (se)	Effect SE (se)	*r* YoP, SE	N
Van Iddekinge et al. 2011, 'Military'	83.7 (39)*	0.0685 (0.0205)	-0.0038 (0.0021)	0.0734 (0.3890)	0.5142 (0.1392)	40
Web et al. 2012, 'C1'	20.7 (16)	0.1092 (0.0593)	-0.0043 (0.0159)	0.1537 (0.7560)	-0.0143 (0.2435)	17
Web et al. 2012, 'D3'	34.5 (16)*	0.4975 (0.1223)	0.0228 (0.0356)	3.2270 (1.8867)	-0.2688 (0.2015)	17
Web et al. 2012, 'D4'	120.5 (42)*	0.3834 (0.0757)	-0.0207 (0.0135)	1.2711 (1.1559)	0.4502 (0.5670)	43
Web et al. 2012, 'R1'	63.4 (23)*	0.2966 (0.0860)	0.0004 (0.0267)	0.7825 (1.2941)	0.5623 (0.1820)	24
Web et al. 2012, 'R2'	21.4 (13)	0.2494 (0.0842)	0.0533 (0.0163)	-0.6593 (1.6734)	-0.2497 (0.2241)	14
Web et al. 2012, 'R3'	64.9 (19)*	0.5133 (0.1099)	-0.0183 (0.0347)	3.2460 (1.9198)	-0.0334 (0.2222)	20
Web et al. 2012, 'S1'	34.9 (28)	0.0826 (0.0488)	-0.0092 (0.0107)	0.2179 (0.8499)	0.02954 (0.1889)	29
Web et al. 2012, 'S3'	21.3 (16)	-0.0702 (0.0834)	0.0166 (0.0273)	1.4289 (0.9780)	0.2171 (0.2609)	17
Web et al. 2012, 'S4'	4.9 (9)	-0.0008 (0.0862)	-0.0065 (0.0242)	0.1701 (1.0515)	-0.4732 (0.1921)	10
Alfieri et al. 2011, 'Children'	40.9 (23)*	0.2125 (0.0619)	0.0053 (0.0044)	0.1762 (0.6911)	0.0818 (0.2114)	24
Alfieri et al. 2011, 'Adolescents'	29.6 (11)*	0.4820 (0.1290)	-0.0040 (0.0082)	-1.3352 (1.1067)	-0.0034 (0.2922)	12
Alfieri et al. 2011, 'Adults'	203.3 (19)*	0.3347 (0.2004)	-0.0372 (0.0157)*	4.3230 (2.1169)	-0.2409 (0.1922)	20
Benish et al. 2011	16.1 (20)	0.3635 (0.0664)	0.0103 (0.0080)	1.1294 (0.8027)	0.4035 (0.2233)	21
Berry et al. 2011	150.0 (20)*	0.7941 (0.1066)	0.1082 (0.0312)	3.4749 (1.9085)	0.3513 (0.2276)	21
Card et al. 2011	142.8 (11)*	0.2258 (0.2080)	0.0100 (0.0221)	-0.8934 (3.0767)	-0.2736 (0.2341)	12
Farber & Doolin, 2011	49.1 (17)*	0.5245 (0.1183)	0.0154 (0.0137)	2.2187 (0.7442)*	-0.0598 (0.2309)	18
Green & Rosenfeld, 2011	19.1 (11)	1.7760 (0.1163)	-0.0192 (0.0194)	4.6925 (2.2448)*	0.2668 (0.3211)	12

Continued on next page

Table 1 – *Continued from previous page*

Meta-Analysis	Q-value (df)	Hedges g (se)	Effect YoP (se)	Effect SE (se)	r YoP, SE	N
Hallion & Ruscio, 2011, 'A'	40.8(23)*	0.1700 (0.0715)	0.0103 (0.0320)	3.2101 (0.8431)*	0.1870 (0.2148)	24
Hallion & Ruscio, 2011, 'I'	19.9 (20)	0.1820 (0.0621)	-0.0717 (0.0346)*	1.7125 (0.9498)	-0.5742 (0.1183)	21
Lucassen et al. 2011	13.8 (15)	0.2377 (0.0542)	-0.0008 (0.0067)	0.6522 (1.1640)	-0.3946 (0.1786)	16
Woodin et al. 2011	41.0 (39)	0.6304 (0.0408)	-0.0006 (0.0058)	0.8939 (0.4975)	-0.4281 (0.1145)	42
Woodley et al. 2011	37.8 (11)*	0.0311 (0.0834)	-0.0146 (0.0125)	-1.9226 (3.5998)	-0.6287 (0.1301)	12