Can meta-analysis be trusted?

Article (Unspecified)

CAN META-ANALYSIS BE TRUSTED?

Andy P. Field

The Psychologist

Dr. Andy P. Field, C.Psychol.
Department of Psychology
University of Sussex
Falmer
Brighton
BN1 9QH
UK

Tel: +44 (0)1273 877150
Fax +44 (0)1273 671320
Email andyf@sussex.ac.uk
http:// www.cogs.susx.ac.uk/users/andyf

Running Head: Meta-analysis

AUTHOR FOOTNOTE

Correspondence concerning this article should be addressed to Andy P. Field, Department of Psychology, University of Sussex, Falmer, Brighton, East Sussex, BN1 9QH. Electronic mail may be sent to andyf@sussex.ac.uk.

ACKNOWLEDGEMENTS

Thanks to Dan Wright, Jeremy Miles, David Clark-Carter and two anonymous reviewers for their comments on an earlier draft.

CAN META-ANALYSIS BE TRUSTED?

Until around 25 years ago the only way to assimilate and evaluate research evidence was through discursive literature reviews in which someone with an interest in a given research topic would accumulate and subjectively evaluate the importance of research findings in that area. These reviews, although informative, are highly reliant on the discretion of the author who, with the best will in the world, could be unaware of important findings or give particular importance to studies that others might believe to be relatively less important (see Wolf, 1986). The failure of literature reviews to provide objective ways to assimilate scientific evidence led scientists to look for a statistical solution. The groundbreaking work of Glass (1976) and Rosenthal and Rubin (1978) paved the way for what we now know as meta-analysis: a statistical technique by which findings from independent studies can be assimilated.

The Use of Meta-Analysis over the Past 20 years

Meta-analysis is generally seen as an important step towards objectifying literature reviews; in fact, the extent to which meta-analysis is now regarded as an accurate and objective way to assimilate research findings is demonstrated by the proliferation of meta-analytic reviews in the major review journals in psychology (e.g. Psychological Bulletin). Figure 1 shows the number of published articles using or discussing meta-analysis that have appeared in peer-reviewed science journals over the last 20 years. As you might expect, during the early 1980s when the technique was being honed by the likes of Hedges and Olkin (1985) and Hunter, Schmidt and Jackson (1982), the discussion of meta-analysis was, to say the least, sparse. Given a few years for the scientific community to absorb these seminal works (and that of Rosenthal, 1991) the use and discussion of meta-analysis suddenly rocketed from under 100 in the late 1980s to the several

hundreds by the early to mid 1990s and to over a thousand by the turn of the century. A substantial proportion of these papers appear in social science and medical journals[1].

Despite the obvious faith that psychologists and other scientists and practitioners have placed in meta-analysis, there is a growing body of evidence to suggest that it is often used incorrectly, and may not be the answer to our literature review prayers after all. This article describes the principles of meta-analysis before reviewing some of the sources of error that might make us doubt whether meta-analysis can be trusted.

Basic Principles of Meta-Analysis

As scientists, we measure effects in samples to allow us to estimate the true size of the effect in a population to which we don't have direct access (Field & Hole, 2003). Imagine I were interested in knowing the effect of sleep-deprivation on people's ability to concentrate (and write articles about meta-analysis!). There is a true effect that sleep deprivation has, but I don't have access to that true effect because it's unlikely that I can sleep-deprive an entire population of people. Instead I use a small sample taken from that population and estimate the true effect that sleep deprivation has, based on the effect in my sample. Now, the chances are that lots of other scientists will also be interested in the effects of sleep deprivation on concentration, and they too will have used samples to estimate the size of the effect that sleep deprivation has. The idea behind meta-analysis is simple: if we take all of these individual studies, quantify the observed effect in a standard way and then combine them, we can get a much more accurate idea of the true effect in which were interested. Effects can be quantified by expressing them as effect sizes (see David Clark-Carter's article in this edition): Cohen (1988) suggested a measure called *d*, but we can use the Pearson correlation coefficient *r*, odds ratios or risk rates. Typically, the choice of measure depends on the conventions of the research

[1] Although it might seem that the use of meta-analysis has accelerated more in science generally than the social sciences, this is likely to reflect the greater number of science journals.

discipline and is not based on statistical reasoning. For example, the correlation coefficient is typically chosen to represent the size of a relationship and Cohen's *d* is used to quantify the degree of difference between group means; however, the correlation coefficient, Pearson's *r*, can be used to quantify differences between means (see Rosenthal, 1991; Field & Hole, 2003). All effect size estimates represent a standardized form of the size of the observed effect, and most can be easily transformed into a different metric and back again (see Rosenthal, 1991); however, there are often statistical reasons for preferring one metric to another.

The first step in meta-analysis is, therefore, to express the effect in each study in a uniform way. So, if we decide to use *r* as our effect size measure, we would need to look at each study and use the data to calculate the value of *r*. The mean of these effect sizes can then be calculated. In addition, although this isn't always the primary concern of meta-analysis, the probability of obtaining that mean can also be computed (see Field, 2000). In short, we can see whether the average effect size is significant, which, to use my sleep deprivation example, would tell us the average size and statistical significance of the relationship between sleep deprivation and concentration across several studies (from which the size of the effect of sleep deprivation on concentration in the population can be inferred).

Meta-analysis can also be used to find out the variability between effect sizes across studies (called tests of the *homogeneity of effect sizes*) and to explain this variability in terms of moderator variables. For example, we might find that in a study that tested concentration using a visual task, the effect of sleep deprivation was larger than in a different study that tested concentration using an audio task. If we had several studies using visual tasks and several using audio ones then we could test whether there was significant variability across effect sizes, but also test whether this variability is caused by the modality of the task (audio vs. visual)—see Field (2003).

Publication Bias and The 'File Drawer' Problem

A major threat to the validity of meta-analysis is that significant findings are more likely to be published than non-significant findings. This is both because researchers may lose interest in non-significant findings and not submit them (Dickersin, Min & Meinert, 1992) and reviewers may scrutinise non-significant findings more closely and reject manuscripts containing them (Hedges, 1984). Rosenthal (1979) calls this the 'file drawer' problem, that is, non-significant research is more likely to end up in the researchers file drawer than in a journal! The extent of this problem should not be underestimated: Sterling (1959) reported that 97% of articles in psychology journals reported significant results, Greenwald (1975) has estimated that significant findings were eight times more likely to be submitted than non-significant ones, and unpublished research can have effect sizes half the value of comparable published research (Shadish, 1992). The effect of this bias is that meta-analytic reviews are likely to over-estimate mean effect sizes (and their significance) because they might not include unpublished studies, in which effect sizes would have been small.

Artefacts

Many different artefacts may also introduce error into a meta-analysis. Although some artefacts may be unique to certain research disciplines, those that stem from the measurement of variables and the general quality of the research apply in all situations. The accuracy of the effect size of a variable will depend largely on how accurately that variable was measured and the error in the measurement of variables is likely to vary across studies. For example, one study might have used a very reliable questionnaire whereas another uses a less reliable one. In addition, correlational research studies can vary in the range of scores elicited from participants (*range variation*), these differences in the range of scores elicited will affect the resulting effect sizes.

More generally, we can say that effect sizes are influenced by the quality of the research. In its simplest form, meta-analysis doesn't take account of the measurement reliability, range

differences, or the general quality of research. Although Hunter and Schmidt (1990) have suggested statistical techniques for correcting for measurement error and range variation, many researchers either do not apply these corrections or apply them incorrectly (Schmidt & Hunter, 1996). Given that artefacts contribute to differences in effect sizes, an alternative approach is to look for significant variability between effect sizes in a meta-analysis (and we've seen earlier that this is possible). If significant variability is found then potential moderator variables can be sought to explain this variability. One possible moderator might be the general quality of the research, and Wolf (1986) suggests that the quality of research can be used as a moderator variable that tests whether the effect size is significantly different in 'well-conducted' and 'badly-conducted' studies. However, this introduces subjective opinion about what is well-conducted research and so perhaps makes the analysis no more objective that the discursive evaluations that meta-analysis seeks to objectify! To make matters worse, although many researchers do test for variability between effect sizes, they rarely act upon the results of these tests (Hunter & Schmidt, 2000).

Misapplications of meta-analysis

Another implication of variability between effect sizes relates to how the meta-analysis is conceptualised and calculated. There are two ways to conceptualise meta-analysis: fixed effects and random effects models[2]. These models differ not only in the theoretical assumptions that underlie them, but also in how the mean effect size and its significance is computed. The fixed-effect model assumes that all studies in a meta-analysis come from a population with a fixed average effect size: studies in the meta-analysis are sampled from a population in which the average effect size is fixed (Hunter & Schmidt, 2000). The alternative assumption is that the average effect size in the population varies randomly from study to study: studies in a meta-

[2] There is actually a mixed model too, but for simplicity I'll ignore it.

analysis come from populations that have different average effect sizes, so, population effect sizes can be thought of as being sampled from a 'superpopulation' (Hedges, 1992).

Statistically speaking, the main difference between fixed- and random-effects models is in the amount of error. In fixed-effects models there is error introduced because of sampling studies from a population of studies. This error exists in random-effects models but in addition there is error created by sampling the populations from a superpopulation. So, calculating the error of the mean effect size in random-effects models involves estimating two error terms, whereas in fixed-effects models there is only one error term.

Another reason why meta-analysis might not be trusted is if the wrong model is used. There is considerable theoretical (Field, 2003; National Research Council, 1992; Hunter & Schmidt, 2000) and empirical (Barrick & Mount, 1991) evidence that real-world data are likely to reflect the random-effects conceptualization (that is, studies come from populations in which the average effect size varies). However, Hedges and Vevea (1998) suggested that the choice of model depends not on the assumptions about the true state of the world, but on the type of inferences that the researcher wishes to make: with fixed-effect models inferences can be drawn only about the studies included in the meta-analysis whereas random-effects models allow inferences that generalise beyond the studies included in the meta-analysis. Psychologists typically wish to generalize beyond the studies included in the meta-analysis and so random-effects models are more appropriate (see Field, 2003; Hunter & Schmidt, 2000).

Despite good evidence that real-world data support a random-effects conceptualisation, the relative simplicity of fixed-effects models has meant that psychologists routinely apply them — even though the random-effects model is more often appropriate. In fact, even when tests reveal significant variability between effect sizes (suggesting a random effects model should be used), psychologists do not act upon these tests and apply fixed-effects models regardless (Hunter & Schmidt, 2000). Hunter and Schmidt (2000) found 21 recent examples of meta-analytic studies using fixed-effects models in Psychological Bulletin (the highest impact review journal for psychology) compared to none using random effects models. What are the

consequences of misapplying fixed-effects models to random-effects data? Well, in short it inflates the estimate of the mean effect size and its significance: normally, using a standard criterion for significance ($p < .05$), we would expect to find a significant average effect size, when there is no effect in the population, in around 5% of cases (the Type I error rate). Hunter and Schmidt (2000) predict that this error rate will increase to between11% and 28% of cases. However, Field (2003) has shown using data simulations that in fact the error rates increase to anywhere between 43% and 80% in certain circumstances. To put this into perspective, of the 21 meta-analyses reported by Hunter and Schmidt (2000), between 9 and 17 of them are likely to have reported significant average effect sizes when, in reality, no significant effect existed within the population.

… and more seriously

A more fundamental issue is that given that we assume real-world data follow a random effects model and effect sizes vary across studies, then what is the value in seeking an average effect size? For example, imagine we tested the efficacy of a powder ('Stat-Whizz') that could magically make you good at statistics. A trial in the USA found an effect size of .45, a replication in Belgium found an effect size of 0, and a further replication in the UK yielded an effect size of -.45. If we assume that these studies had equal sample sizes and so were equally weighted in the meta-analysis, then the resulting average effect size would be 0—there would be a non-significant effect. Readers of such a meta-analysis might conclude, therefore, that Stat-Whizz was an ineffective drug. Of course, this conclusion is wrong: the drug worked in the USA, didn't work in Belgium and had a negative effect in the UK. As such, the issue of interest is not so much the overall effect of the drug, but at what levels the drug works: the fact that the drug doesn't work on the English is of little interest to all of the Americans for whom the drug is effective! A retort to this is that such variability would be picked up by tests of the homogeneity of effect sizes. However, as mentioned previously, researchers frequently fail to act upon such tests (Hunter & Schmidt, 2000). One implication of these observations is that moderator

analysis, in which we look for possible variables that explain the variation between effect sizes, may be more useful than looking at average effect sizes.

Methods of meta-analysis

A final possible source of error in meta-analysis could be due to problems inherent in the method used. Three methods of meta-analysis have been popular: the methods devised by Hedges and colleagues (Hedges & Olkin, 1985; Hedges, 1992; Hedges & Vevea, 1998), Rosenthal and Rubin's method (1978), or that of Hunter and Schmidt (1990). Hedges and colleagues have developed both fixed- and random-effects models for combining effect sizes, Rosenthal and Rubin have developed only a fixed-effects model, whereas Hunter and Schmidt label their method a random-effects model. The computations of these various methods differ, and the technical details of these differences are well-documented elsewhere and are beyond the scope of this review (see Field, 2001).

Several recent studies have compared these methods. Johnson, Mullen and Salas (1995) compared the Hedges-Olkin (fixed-effect), Rosenthal-Rubin and Hunter-Schmidt meta-analytic methods by manipulating a single data set. They concluded that the significance of the mean effect size differed substantially across the methods: the Hunter and Schmidt method reached more conservative estimates of significance than the other two methods so should be used cautiously. Schmidt and Hunter (1999) subsequently claimed that Johnson et al. incorrectly applied their method and showed that, theoretically, when the method was correctly applied, their method was comparable to that of Hedges. Field (2001) highlighted some other concerns with Johnson et al.'s methods and rectified these concerns in a series of simulations that compared the methods across a variety of situations. Field found that when comparing random-effects methods, the Hunter-Schmidt method yielded the most accurate estimates of population effect size across a variety of situations. However, neither method controlled the Type I error rate when 15 or fewer studies were included in the meta-analysis, and that the method described by Hedges and Vevea (1998) controlled the Type I error rate better than the Hunter-Schmidt method when 20 or more studies were included. In a more recent set of simulations,

Field (2002) demonstrated that across a far-ranging set of situations both methods produce biased estimates of the population effect size: however, the biases in the Hunter-Schmidt method are not as large as in Hedges' method. Hedges method did tend to keep tighter control of the Type I error rate but with 80 or more studies in the meta-analysis, there was little to separate the two methods. With fewer studies in the meta-analysis (20-40), Hedges method controlled the Type I error rate considerably better than Hunter and Schmidt's method. As a general rule, neither method was accurate when fewer than 20 studies were in the meta-analysis.

Can meta-analysis be trusted?

To sum up, meta-analysis has come to be seen as the saviour of the literature review, but perhaps unjustly. As with all statistical procedures, the results are only as good as the data available and the person performing the test: if fixed-effects models continue to be routinely applied to psychological data then we risk finding inflated effects. In terms of which method to apply, if the primary interest is in estimating the effect in the population then what matters is whether it's better, in the context of the question you're trying to address, to underestimate (Hunter-Schmidt) or overestimate (Hedges) the effect size in the population. If the significance of this estimate is important, and there are 80 or more studies in the meta-analysis then either method will be fairly reliable, but with 20-40 studies Hedges' method is preferable, and significance tests should not be conducted at all with fewer than 20 studies in the meta-analysis.

Of course, there is more to meta-analysis than this simple summary suggests. I've hinted at the fact that moderator variables may often be more interesting than the average effect size. Also researchers have to give greater consideration to controlling for other sources of error because small statistical differences between the methods discussed here may be relatively unimportant compared to biases from other sources such as using unreliable measures.

Word Count: 3058

REFERENCES

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, *44*, 1-26.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd Ed.). New York: Academic Press.

Dickersin, K., Min, Y.-I., & Meinert, C. L. (1992). Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *Journal of the American Medical Association*, *267*, 374–378.

Field, A. P. (2000). *Discovering statistics using SPSS for Windows: advanced techniques for the beginner*. London: Sage.

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*, 161–180.

Field, A. P. (2003). The problems in using Fixed-effects models of meta-analysis on real-world data. *Understanding Statistics, in press*.

Field, A. P. (2002). Is the meta-analysis of correlation coefficients accurate when population effect sizes vary? *Manuscript submitted for publication*.

Field, A. P., & Hole, G. (2003). *How to design and report experiments*. London: Sage.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5I*, 3-8.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Hedges, L. V. (1984). Estimation of effect size under non-random sampling: the effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*, 61–85.

Hedges, L. V. (1992). Meta-Analysis. *Journal of Educational Statistics*, *17*, 279-296.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V. & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486-504.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of Meta-analysis: correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: implications for cumulative knowledge in Psychology. *International Journal of Selection and Assessment*, *8*, 275–292.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: cumulating research findings across studies*. Beverly Hills, CA: Sage.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology*, *80*, 94-106.

National Research Council (1992). *Combining information: Statistical issues and opportunities for research*. Washington, D.C.: National Academy Press.

Rosenthal, R. (1979). The 'file drawer' problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (revised). Newbury Park, CA: Sage.

Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: the first 345 studies. *Behavior and Brain Sciences*, *3*, 377-415.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: lessons from 26 research scenarios. *Psychological Methods*, *1*, 199-223.

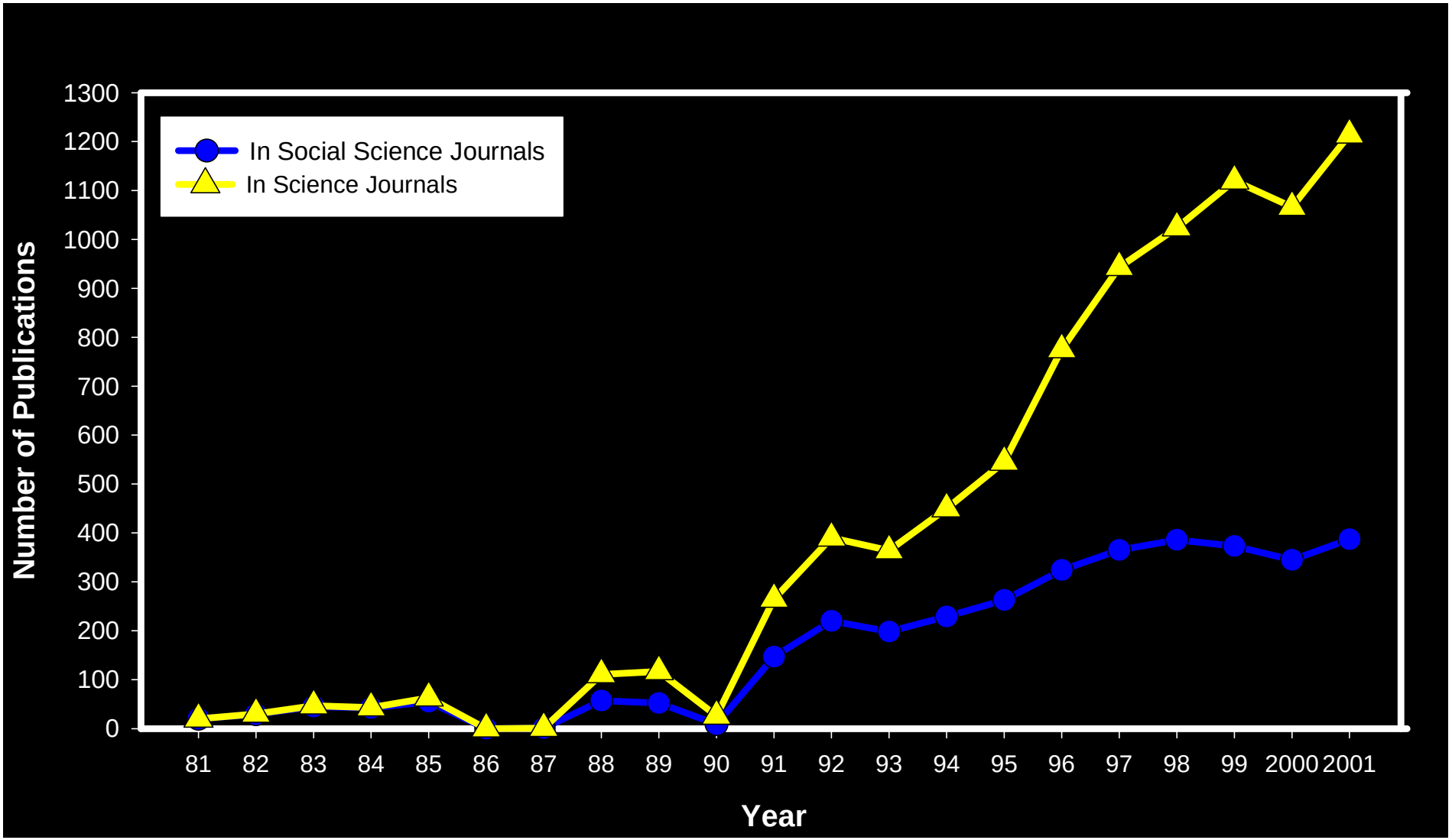
Schmidt, F. L., & Hunter, J. E. (1999). Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen, and Salas (1995). *Journal of Applied Psychology*, *84* (1), 144-148.

Shadish, W. R. (1992). Do family and marital psychotherapies change what people do? A metaanalysis of behavioural outcomes. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis & F. Mosteller (eds.), *Meta-Analysis for Explanation: A Casebook* (pp.129-208). New York: Sage.

Sterling, T. C. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

Wolf, F. M. (1986). *Meta-Analysis*. Sage university paper series on quantitative applications in the social sciences, 07–061. Newbury Park, CA: Sage.

FIGURES


In Social Science Journals
In Science Journals
Number of Publications
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
2000
2001
Year