Meta-analysis and Its Applications in Horticultural Science

Ingram Olkin

Department of Statistics, Stanford University, Stanford, CA 94305-4065

Douglas V. Shaw

Pomology Department, University of California, Davis, CA 95616

Horticulture is a vastly heterogeneous subject; therefore, accurate estimates of the magnitude of research effort on topics directly or tangentially related to horticulture are difficult to obtain. A few examples will suffice for demonstration:

- A search of the Univ. of California library system general catalogue identified 434 periodicals and symposium volumes cross-referenced under the subject word "horticulture" within the past 10 years. This list included very few journals dedicated to the specific disciplines that support horticultural research, such as genetics or plant physiology.
- The International Society for Horticultural Science (ISHS) 1986 directory lists 16,650 members at 1250 institutes in 68 countries (ISHS, 1986).
- Domestically, the U.S. Dept. of Agriculture (1991) listed 5338 project titles, managed by 1849 full-time equivalents, involving state, federal, and private-sector horticultural research. The annual research support for these projects is 419 million U.S. dollars, and the farm-gate (primary product) value for fruit, vegetable, and ornamental products supported by this research exceeds 32 billion dollars per year nationwide.

By any measure, the quantity of information generated each year, the effort invested, and the value of the horticultural products at which the research support is aimed is staggering.

Within more narrow disciplines, research usually is conducted by several investigators working on similar questions, and often generates substantial information. For example, a search of the National Agricultural Library (AGRICOLA) records identified 363 references to research dealing with strawberries in 1992 alone. Fifty-one of these references in 25 sources were specific to strawberry disease incidence and control.

One of the most important tasks confronting the horticultural scientist is the assimilation and evaluation of research results pertinent to their subject area. Interpreting the

results of research relating to a specific topic is a critical step, both in evaluating what is currently known and in formulating hypotheses for further testing and study. Although obtaining a thorough knowledge of the literature is universally accepted as good scientific practice, the process of compiling information is rarely formalized.

One obstacle to the traditional literature review is that published results from similar or identical experiments frequently disagree. Although physical and biotic factors may explain heterogeneous results, even well-designed studies are often limited in statistical power. As a further complicating factor, the pertinent questions in horticulture often concern multiple factors, each with several levels. In most cases it is impossible to test all relevant factors in a single study without compromising the amount of experimental effort that can be dedicated to replication of individual treatments. Despite the vast resources and human effort dedicated to horticultural research, the resources available for individual projects limit the scope and precision of experimental results.

When results do not agree, it is important to know whether real differences exist, or whether the inconsistency is the consequence of experimental error. Even when general trends are discernible, inference about the magnitude of an effect may be far more important than a simple statistical statement. Traditional treatments of the literature are of little use in resolving such issues. Furthermore, when substantial information has been generated about a topic, the value added by new studies may represent little more than additional data points.

The problems inherent in the analysis of scientific experimentation are not limited to horticultural science. Foreexample, meteorologists have grappled with the questions regarding the effectiveness of weather modification. The following is a quote from Todd (1986):

"Some people are hoping for the day when an elegantly designed experiment, brilliantly evaluated, will make precipitation management credible to those who look at the field from the outside, but it doesn't appear as though credibility will come that way. Rather, it will come from reporting on information from a large number of imperfect experiments and operations, that show even though there is diversity in the way different types of clouds respond to seeding, similar clouds respond similarly around the world."

In medicine, studies are being conducted at an ever increasing rate. Providing guidance to the clinician as to the effectiveness of inter-

ventions is essential. Consequently, the number of meta-analyses, overviews, or syntheses has soared exponentially. A recent article in the *New York Times* (Jan. 1994) on the integration of medical research offers a definition that captures the essence of meta-analysis:

"A meta-analysis aims at gleaned more information from existing data by pooling the results of many smaller studies and applying one or more statistical techniques. The benefits or hazards that might not be detected in small studies can be found in a meta-analysis that uses data from thousands of patients."

Meta-analysis, therefore, is an expression used to identify a set of statistical techniques developed to provide an objective comparison of research results from independent studies. It differs from more classical methods of verbal/subjective synthesis primarily in that it depends on quantification for its synthesis. These techniques offer a coherent method for combining research results across studies that can provide insight unlikely to be derived from the individual studies themselves. Other components of the formalism allow reduction of bias, notably the detection of publication bias toward significant results, and increased ability to resolve conflicting results by identifying potential explanatory variables.

The idea of combining research results from independent studies dates to the early parts of this century (for example, Cochran, 1937, 1938; Fisher, 1932; Pearson, 1950). Although the early motivation for combining the results of independent studies was obtained from agricultural experiments, the methods of modern meta-analysis were developed for use in the social and behavioral sciences, where limits to experimental control can generate frequent conflict in experimental results. However, the potential advantages of a statistical combination, namely improved statistical power and robustness, have been widely applied, notably in the field of medical research and applications to public policy. The objective of this article is to present and demonstrate some of the methods used in meta-analysis that might be useful in horticultural science.

APPROACH

Two literature sets were chosen for demonstration, one concerning biological control and one concerning chemical control of an important strawberry pest, *Tetranychus urticae* Koch, in California. These examples were chosen because, together, they illustrate several kinds of information accessible through

Received for publication 13 Jan. 1995. Accepted for publication 24 May 1995. We are grateful to Carolyn Pickel, John Trumble, Douglas Walsh, Norman Welch, and Frank Zalom for providing raw data or primary unpublished results. We also thank Eddie Hak-sing Ip for his help with some of the computations. This work was supported in part by the National Science Foundation. The cost of publishing this paper was defrayed in part by the payment of page charges. Under postal regulations, this paper therefore must be hereby marked advertisement solely to indicate this fact.

meta-analysis. First, similar research has been conducted over many years, and the results (together with the professional opinion of scientists and growers) are inconclusive. Second, the relative value of these pest control alternatives has not been resolved, and, due to design limitations, they rarely have been tested in combined studies. Third, a practical consideration is that much of the published work does not include the information required for performing even the simplest of meta-analyses (an issue to be discussed later), and one of us (D.V.S.) had professional relationships that allowed access to the primary results needed.

Biological control was summarized using 10 studies conducted between 1964 and 1991 regarding the effect of *Phytoseiulus persimilis* Athias-Henriot, an introduced predatory mite, on *T. urticae* (Table 1). We restricted our evaluation to studies that compared strawberry yield for treated and nontreated replicated plots. In three cases, primary data were obtained from the authors, and three studies that included yield summaries in their text were omitted because primary results or information about treatment standard deviations were not available. The studies included here were conducted using a broad range of cultivars, planting systems, and predator release rates. A much larger body of literature is available that compares the predator-prey dynamics, but this study set comprises a rather exhaustive search for studies that consider the consequences of biological control for strawberry yield in a Mediterranean climate.

Chemical control was summarized using 10 studies conducted between 1966 and 1991 (Table 2). These experiments were conducted either to test the effectiveness of various compounds for controlling *T. urticae*, or to compare varieties for pest tolerance in California. We limited our evaluation to registered insecticides, and to studies that included replicated trials in which yield was measured. In six cases, primary data were obtained from the authors (one of which was unpublished in any form). We omitted two studies that included yield summaries in their text, because primary results or information about treatment standard deviations were not available. As in the case of the biological control study set listed above, the studies summarized here were conducted using a broad range of cultivars and planting systems.

Two caveats regarding the chemical control study set are important to note. First, a much larger body of literature was available than could be included, although studies that included yield data were not readily available. These studies were chosen with the intent of providing a manageable demonstration and constituted a random sample of those studies most easily retrieved. Second, experiments conducted to test the effectiveness of individual insecticides may not reflect the level of control achieved by more comprehensive programs used by growers. This latter issue concerns the validity of the control group, and will be discussed further.

These data sets exhibit several general prin-

ciples concerning a search of the literature, namely, that the search should be exhaustive. If only studies with significant results are published, then we are including a segment of studies from one side of the distribution, which will result in a biased estimate of the total treatment effect. Because the results of many horticultural experiments are reported in working papers, abstracts, or technical reports, the search should not rely only on computer literature data bases or journals. The demonstration data sets also note the deletion of several studies for a variety of reasons. The exclusion criterion should always be stated explicitly, and an independent corroboration for the deletion of studies will help remove a potential selection bias.

STATISTICAL METHODS AND DEMONSTRATIVE RESULTS

Meta-analysis includes procedures for synthesis of results from a diversity of experimental designs, but the demonstration here will concentrate on three topics: combining tests of significance across studies, judging the relative size of the difference between treatment and control over many studies, and computing confidence intervals for these combined treatment effects. For completeness, data are provided for biological and chemical control sets. However, for simplicity of presentation, methodology will be illustrated using data for biological control only.

Summary of individual studies. The results

Table 1. Compiled descriptive information for 10 studies of biological control of spider mites in strawberries.

Reference	Study no.	Test year	Cultural system	Release rate ^a	Comments
Oatman and McMurtry, 1966	I	1965	Summer	160,000/acre	Shasta, yield in kg/plot, nonsignificant pruning treatments combined
Oatman et al., 1966	II	1965	Summer	320,000/acre	Shasta, yield in kg/plot, nonsignificant pruning treatments combined
Oatman et al., 1967	III	1966	Summer	384,000/acre	Shasta, yield in g/plant, older leaves removed
	IV	1966	Summer	384,000/acre	Shasta, yield in g/plant, older leaves not removed
Oatman et al., 1976	V	1975	Summer	250,000/acre	Tufts, yield in g/plant
	VI	1975	Summer	125,000/acre	Tufts, yield in g/plant
Oatman et al., 1977	VII	1976	Summer	224,000/acre	Tufts, yield in g/plant
Pickel et al., 1988	VIII	1988	Summer	>>200,000/acre	Pajaro, yield in g/plant
Trumble and Morse, 1993	IX	1989	Winter	40,000/acre	Selva, yield in lbs/acre ^a
	X	1991	Winter	40,000/acre	Selva, yield in lbs/acre ^a

^aRelease rates and yields are provided in British units to facilitate comparison with the original data.

Table 2. Compiled descriptive information for 10 studies of chemical control of spider mites in strawberries.

Reference	Study no.	Test year	Cultural system	Chemical	Comments
Oatman et al., 1967	XI	1966	Summer	Binapacryl	Shasta, yield in g/plant, chemical trial
Oatman et al., 1975	XII	1974	Summer	Phorate	Shasta, yield in g/plant, chemical trial
Pickel et al., 1988	XIII	1988	Summer	Commercial combination	Pajaro, yield in g/plant, yield trial
	XIV	1988	Winter	Commercial combination	Selva, yield in g/plant, plants chilled for 3 weeks, yield trial
Pickel et al., 1989	XV	1989	Winter	Commercial combination	12 Cultivars and selections, yield in g/plant, plants chilled according to commercial standard, genetic tolerance trial
Zalom et al., 1991 ^a	XVI	1991	Winter	Commercial combination	12 Cultivars and selections, yield in g/plant, plants chilled according to commercial standard, genetic tolerance trial
Walsh et al., 1992a ^a	XVII	1992	Winter	Commercial combination	12 Cultivars and selections, yield in g/plant, plants chilled according to commercial standard, genetic tolerance trial
Walsh et al., 1992b ^a	XVIII	1992	Winter	Commercial combination	Selva, yield in g/plant, plants chilled for 3 weeks, yield trial with mite inoculation
Trumble and Morse, 1993	XIX	1989	Winter	Avid/four applications	Selva, yield in lbs/acre ^a , plants chilled for 2 weeks, mite control trial
	XX	1991	Winter	Avid/four applications	Selva, yield in lbs/acre ^a , plants chilled for 2 weeks, mite control trial

^aUnpublished results.

^aYields are provided in British units to facilitate comparison with the original data.

of individual studies and one-tailed probability values are given for the biological (Table 3) and chemical (Table 4) control studies. One-tailed probabilities (p) for the former set range from 0.034 to 0.482. Note that conventional comparisons with Student's *t* use a two-tailed comparison, which makes no assumption about the direction of treatment response. Thus, although all of the biological control effects are positive, none of the cases provide evidence for a statistically significant effect of the predator treatment at the 5% level, using standard comparisons. Similar results for the chemical control set yield one-tailed probabilities (p) from 0.001 to 0.189 (Table 4). All of the chemical treatment effects were positive, and six of the 10 studies would give statistically significant differences between treated and control groups at the 5% level when compared using a two-tailed test.

Combining tests of significance: Cumulative probabilities. One of the simplest methods for combining significance tests was developed by Fisher (1932). In this method, Fisher showed that under the null hypothesis the statistic

$$-2 \sum \log p_i = -2 \log \pi$$
 [1]

has a χ^2 distribution with $2k$ degrees of freedom. In Eq. [1], the p_i are the probabilities of finding a larger value for the treated group than for the control group in each of the k studies, given that there is no real difference between treated and control sets: the individual p_i are nothing more than the probabilities, or p values, obtained from a one-tailed probability test. In the present context, the specific test is Student's two-sample *t* test. However, in other contexts, a different test might be used. The value calculated in Eq. [1] can then be compared with a standard χ^2 table with $2k$ degrees of freedom to obtain a point significance level for the treatment over all k studies.

With six of 10 studies showing strong significant differences between treatment and control groups for the chemical set, most horticulturists would anticipate the presence of a real effect. However, the interpretation for the biological control set is less clear: none of the individual studies provide compelling evidence for a significant treatment effect, yet all of the estimated effect sizes are positive. Using Eq. [1] to infer a combined significance value provides a resolution to this dilemma: the cumulative χ^2 for the biological control studies is 36.48, which for $2k = 20$ degrees of freedom gives an associated p value of 0.013. Thus, considering all cases, there is evidence of a significant treatment effect at the 5% level of significance. For comparison, the cumulative $\chi^2_{20} = 94.46$ for the chemical control set, with an associated p value of $P < 0.001$.

Although p values are not used to produce confidence intervals, they are useful for diagnostics. In any series of experiments it may be valid to question the extent to which those studies included in a meta-analysis are heterogeneous due to unspecified explanatory variables, or are affected by outliers. We now describe two sensitivity analyses that might be used to evaluate homogeneity of results, and

Table 3. Compiled results for individual biological control studies in strawberries.

Study no.	Reps	Group				Percent increase ^a	t ^b	p ^c	d ^d
		Treated		Control					
		Mean yield	SD	Mean yield	SD				
I	4 ^a	134.3	2.9	129.8	15.0	3.4	0.657	0.274	0.57
II	4 ^a	130.5	17.3	129.8	15.0	0.6	0.048	0.482	0.04
III	4	268.3	60.5	212.3	43.2	26.4	1.507	0.091	1.07
IV	4	295.0	74.0	254.8	81.4	15.7	0.731	0.246	0.52
V	5	842.1	81.6	729.4	91.7	15.5	2.053	0.037	1.30
VI	5	767.4	104.6	729.4	91.7	5.2	0.611	0.279	0.39
VII	4	471.1	375.9	279.5	289.3	68.5	0.808	0.225	0.57
VIII	4	1,035	150.0	855	59.9	21.1	2.229	0.034	1.58
IX	4	11,787	2,487	10,603	3,648	11.2	0.536	0.306	0.38
X	4	13,661	6,152	10,202	2,454	33.9	1.044	0.168	0.74

^aPercent increase in mean yield of treated over control groups.

^bSee text for detailed descriptions of *t*, *p*, and *d*.

^cFour replications for treatments, two for the control; other studies have a common number of replication for treated and control groups.

Table 4. Compiled results for individual chemical control studies in strawberries.

Study no.	Reps	Group							
		Treated		Control		Percent increase ^c	t ^b	p ^b	d ^b
		Mean yield	SD	Mean yield	SD				
XI	4	322.8	46.7	254.8	81.4	26.7	1.449	0.099	1.03
XII	4	172.8	53.0	139.8	45.4	23.6	0.946	0.189	0.67
XIII	4	1,107	250.1	854.5	59.6	29.5	1.964	0.049	1.39
XIV	4	1,171	60.8	967	64.4	21.1	4.607	0.002	3.26
XV	36 ^a	1,992	104.5	1,283	132.7	55.3	25.186	0.000 ^w	5.94
XVI	24 ^a	1,566	64.7	1,340	148.9	16.8	6.820	0.000 ^w	1.97
XVII	24 ^a	1,679	88.7	1,036	75.4	62.1	27.059	0.000 ^w	7.81
XVIII	3	1,115	143.4	446	58.0	150.0	7.491	0.001	6.12
XIX	4	13,040	2,677	10,603	3,648	23.0	1.077	0.160	0.76
XX	4	21,717	6,770	10,202	2,454	112.9	3.198	0.010	2.26

^aPercent increase in yield of treated vs. control groups.

^bSee text for detailed descriptions of *t*, *p*, and *d*.

^cStandard deviations were calculated using pooled within-cultivar sums of squares to eliminate genetic main effects.

^d*P* values <0.00005 were rounded to 0.0001 for further calculations.

illustrate these using the biological control example.

The 10 one-tailed p values for biological control results in ascending order are: 0.034, 0.037, 0.091, 0.168, 0.229, 0.246, 0.274, 0.279, 0.306, and 0.482, which give an overall p value of 0.013 (using the Fisher procedure in Eq. [1]). That these values comprise a homogeneous set of results can be ascertained by using the range as a test statistic. Under the null hypothesis, these k values come from a uniform distribution. The critical value for a range of p values over the k studies can be obtained by comparing the observed range with a critical value obtained from the beta distribution with $k-1$ and 2 degrees of freedom (Beyer, 1966). In our case, the critical value is 0.60, which is considerably larger than the actual range (0.482–0.034 = 0.448). Thus, the range of p values obtained for this study set appears consistent with that expected from repeated sampling from a homogeneous set of experiments.

A second diagnostic is to delete one study at a time and to compute an overall p value using the remaining nine studies. This exercise for the biological control studies gives the results summarized in Table 5. The range of cumulative p values is 0.009 to 0.404, which is exceedingly homogeneous, and does not indicate the presence of a particularly influential

study. When heterogeneity is detected, this diagnostic indicates which studies need to be examined in greater detail.

Another useful diagnostic is to delete 5% to 10% of studies with the highest level of significance. This procedure has the effect of being ultraconservative in drawing the conclusion that treatment effects are real. Because the procedure is carried out after seeing the data (post hoc), its application requires a more sophisticated analysis than can be described here. We refer the reader to Olkin and Saner (1993), but note that when the smallest p value from Table 3 (0.034) was omitted, an overall

Table 5. Cumulative χ^2 values and probabilities for biological control studies in strawberries, with individual studies deleted one at a time.

Study omitted	χ^2 (18 df)	Cumulative probability ^a
I	33.9	0.0130
II	35.0	0.0094
III	31.7	0.0239
IV	33.7	0.0138
V	29.9	0.0385
VI	33.9	0.0128
VII	33.5	0.0145
VIII	29.7	0.0404
IX	34.1	0.0122
X	32.9	0.0171

^a*P* value obtained with Eq. [1] when each specified study is omitted from the cumulative result.

p value of 0.040 was obtained, and the overall significance level changed little. This diagnostic is especially useful in judging the confidence given to the overall result.

The Fisher method given by Eq. [1] is not the only method available for combining p values. Another commonly used procedure is that of Stouffer (Stouffer et al., 1949) in which each p value is translated into a z-score. If there are k studies, z is the mean of their z-scores, then $\sqrt{k}z$ has a standard normal distribution, from which an overall p value can be obtained.

Comparing differences between treatments. In any single study, several options are available for comparing treated and control group means. In our examples, raw yield values or proportionate increases due to biological or chemical control treatments are possibilities. When synthesizing research results, these methods suffer in that the differences assessed using these variables depend on the scale of measurement and on the error rate of the study itself. One alternative is to use standardized sample effect sizes, d, for comparison, calculated as:

$$d = \frac{X_T - X_C}{s} \tag{2}$$

where $\bar{X}_T$ and $\bar{X}_C$ are means for the treated and control groups, respectively, and s is the pooled standard deviation. The value d is a biased estimate of the population effect size $\delta = (\mu_T - \mu_C)/\sigma$, where μ_T and μ_C are the true treatment and control means and σ is the common standard deviation; the biases are small when sample sizes are large and similar among experiments. More specifically, $[(N - 3)/(N - 2.25)]d$, where $N = n^T + n^C$, is approximately unbiased and shows what corrections are needed for small sample sizes. (For a more detailed discussion of bias, see Hedges and Olkin, 1985). We note that the estimated effect sizes (d) range from 0.04 to 1.58 for biological control cases (Table 3) and from 0.67 to 7.81 for chemical control cases (Table 4). Although the number of replications is small in nearly all studies, for the sake of simplicity, we have not invoked corrections for bias; the unbiased estimates are somewhat smaller in all cases. A number of estimates, such as maximum likelihood, Bayes, and Shrinkage, for the population effect size appear in the literature; we use the notation "d" as a generic estimate. A description of alternative estimates and their properties is provided in Hedges and Olkin (1985).

To obtain a composite estimate for a given treatment effect size over a number of studies, those from individual studies must be weighted. Large sample theory indicates that the estimated variance, $\hat{\sigma}^2(d_i)$, of the ith study effect size can be calculated as:

$$\hat{\sigma}^2(d_i) = \frac{n_i^T + n_i^C}{n_i^T n_i^C} + \frac{d_i^2}{2(n_i^T + n_i^C)} \tag{3}$$

In Eq. [3], n_i^T and n_i^C are the numbers of experimental units (observations or replications) for treated and control groups, respectively, and d,

is the estimated effect size for the ith study, i = 1...k. The weight, w_i , for each effect size estimate is inversely proportional to its variance. Thus, effect sizes estimated from studies with small variability are given larger weights. That is, greater weight is given to estimates in which we have greater confidence. The weights are computed as:

$$w_i = \frac{1/\hat{\sigma}^2(d_i)}{\sum [1/\hat{\sigma}^2(d_i)]} \tag{4}$$

and the combined estimate, d_+ , of effect size is calculated as:

$$d_+ = w_1 d_1 + w_2 d_2 + \dots w_k d_k = \sum w_i d_i \tag{5}$$

The estimated variance, $\hat{\sigma}^2(d_+)$, of the combined effect size estimate is calculated as:

$$\hat{\sigma}^2(d_+) = \frac{1}{\sum [1/\hat{\sigma}^2(d_i)]} \tag{6}$$

and a confidence interval for the combined effect size is determined as $d_+ \pm c_{\alpha/2} \hat{\sigma}_{d_+}(d_+)$, with $c_{\alpha/2}$ being the appropriate coefficient obtained from the normal distribution for the chosen confidence interval (e.g., 1.645 or 1.96 for a 90% or 95% confidence interval, respectively).

Tables 6 and 7 provide the details required for the application of Eqs. [3], [4], and [5] to obtain an overall effect size estimate and its associated confidence interval. Using these equations, cumulative effect sizes were com-

puted as 0.714 and 2.714, and 95% confidence intervals for the population effect size δ determined using Eq. [6] were $0.253 < \delta < 1.174$ and $2.299 < \delta < 3.129$ for biological and chemical control examples, respectively. It is customary in some fields to provide a diagram that shows the individual-study confidence intervals together with an overall confidence interval (see Fig. 1 for an example using the biological control studies). This type of display provides a good graphic illustration of the effect in each study. Confidence intervals for each study were obtained in this case from the two-sample t statistic.

The above example makes clear that, if a real treatment effect is present, it will eventually show statistical significance, given that sufficiently replicated trials were conducted. A far more important consideration in most horticultural studies (not to mention policy debates) should be the actual size of the treatment effect, and the degree of confidence we have in this estimated effect size. In this respect, the comparisons using Fig. 1 or Tables 6 and 7 are most illuminating. Two important conclusions could be derived at this point that are not apparent from consideration of treatment significance levels. First, to the extent that these experiments are a representative sample of the differing treatments, chemical control is better at increasing strawberry yield by a factor of 3.81. Second, note that the 95%

Table 6. Weighted effect sizes and confidence intervals for biological control studies in strawberries.

Study no.	Effect size (d _i)	Study confidence intervals	$\hat{\sigma}^2(d_i)^*$	$1/\hat{\sigma}^2(d_i)^z$	w_i^*	$w_i d_i^*$
I	0.57	-1.02, 2.29	0.777	1.287	0.071	0.040
II	0.04	-1.66, 1.74	0.750	1.333	0.074	0.003
III	1.07	-0.43, 2.55	0.571	1.751	0.097	0.104
IV	0.52	-0.89, 1.93	0.517	1.934	0.107	0.056
V	1.30	-0.06, 2.66	0.484	2.066	0.114	0.148
VI	0.39	-0.97, 1.75	0.407	2.457	0.136	0.053
VII	0.57	-0.85, 1.98	0.520	1.923	0.106	0.060
VIII	1.58	-0.01, 3.16	0.655	1.527	0.084	0.133
IX	0.38	-1.02, 1.78	0.509	1.965	0.108	0.041
X	0.74	-0.69, 2.17	0.534	1.873	0.103	0.076
Totals				18.116	1.000	0.714
95% Confidence interval					0.253, 1.174	

*Values for $\hat{\sigma}^2(d_i)$ were obtained using Eq. [3], values w_i were obtained using values of $1/\hat{\sigma}^2(d_i)$ divided by its column total, d_i is given by the column total for $w_i d_i$.

Table 7. Weighted effect sizes and confidence intervals for chemical control studies.

Study no.	Effect size (d _i)	Study confidence intervals	$\hat{\sigma}^2(d_i)^*$	$1/\hat{\sigma}^2(d_i)^z$	w_i^*	$w_i d_i^*$
XI	1.03	-0.44, 2.50	0.566	1.767	0.079	0.081
XII	0.67	-0.36, 1.70	0.528	1.894	0.085	0.057
XIII	1.39	-0.15, 2.93	0.621	1.610	0.072	0.100
XIV	3.26	1.15, 4.47	1.163	0.860	0.039	0.127
XV	5.94	4.86, 7.01	0.300	3.333	0.150	0.891
XVI	1.97	1.28, 2.66	0.124	8.065	0.362	0.713
XVII	7.81	6.14, 9.47	0.719	1.391	0.062	0.484
XVIII	6.11	2.30, 9.92	3.784	0.264	0.012	0.073
XIX	0.76	-0.67, 2.19	0.536	1.866	0.084	0.064
XX	2.26	0.49, 4.03	0.820	1.220	0.055	0.124
Totals				22.270	1.000	2.714
95% Confidence interval					2.299, 3.129	

*Values for $\hat{\sigma}^2(d_i)$ were obtained using Eq. [3], values w_i were obtained using values of $1/\hat{\sigma}^2(d_i)$ divided by its column total, d_i is given by the column total for $w_i d_i$.

confidence limits for the two estimates of effect size do not overlap; thus, the two methods of control differ significantly.

Subgroup comparisons: Application rates and cultural systems. Further analysis of the variation can proceed by a systematic investigation of potential explanatory variables. The intent here is to separate sampling error from other systematic sources of difference among the experimental results. Table 8 presents a summary of cumulative effect sizes and confidence intervals for several extended comparisons within our demonstrative results. For this demonstration, we have compared results with trials established using differing cultural systems (summer vs. winter planting) and also divided the cases into subsets with high and low rates of application for the biological and chemical control groups.

Several insights emerge from these comparisons. First, as suggested by the sensitivity analyses performed using p values, the biological control series is remarkably homogeneous: all of the estimated effect sizes have overlapping 95% confidence intervals. Conversely, the chemical control set contains some heterogeneous subsets. Chemical control was significantly more effective at increasing yield

in winter plantings than in summer plantings, and there are significant differences between occasional application of pesticide and application at levels considered effective in commercial practice. Any number of pairwise comparisons might be made from this type of subdivision, although caution is needed for interpretation. For example, only three summer-planted experiments are available within the chemical control group, and just one of these was conducted using a commercial combination of chemical treatments.

Obtaining additional information. Another informative way to treat a series of experiments is to compile cumulative effect size estimates and confidence intervals in historical order. This type of analysis can generate two kinds of information. First, trends in the effectiveness of a treatment over time may become apparent, which in turn may suggest additional explanatory variables, such as the manner in which the treatments are administered. In our example, there were no apparent trends for biological control, whereas there were indications of increased chemical effectiveness with increases in time. Second, this treatment helps visualize the added information derived from subsequent studies. In other

words: Will similar studies contribute to our confidence in the general trend meaningfully, and, if not, what types of studies will be required?

Other applications of meta-analysis are available for synthesizing research results for correlations, analyses of variance, and regressions. This paper provides analyses using only yield as a measure. However, in most experiments, there will be several correlated measures. The appropriate analysis will involve a discussion of multivariate procedures [see Hedges and Olkin, 1985 (Chapter 10); Gleser and Olkin, 1993 (Chapter 22)]. For single measures, combined p values can follow the Fisher method presented here, using the statistical test appropriate to each method of analysis. Weighting individual correlation or regression coefficients to obtain combined overall estimates is also possible and follows procedures analogous to those presented for combining effect sizes.

DISCUSSION

In any research synthesis, it is critical that the performance of the control set accurately reflect the control population, as disparate results may be explained by an unusual control result rather than a failure of the treatment itself. The consequences of an unusual control generally fall into two categories which depend on the choice of a control treatment. First, when performance of the control is expected to be less than that of the treated group, an unusually successful control performance will reduce the opportunity for detecting significant treatment effects and will result in standardized effect sizes that are underestimates. An example here is a test for the effectiveness of a drug in preventing mortality due to a specific disease: no effect will be detected if none of the control group become ill. An analogous example from our demonstration results is that neither biological nor chemical control will seem effective for increasing strawberry yield when plantations are free of the target pest.

An alternative concern is critical when the control is expected to be superior to the treated set, but fails due to factors extraneous to the experiment. For example, imagine that we had compared biological and chemical treatments to reduce spider mites and increase yield in strawberry, with chemical treatment chosen as the control. Excessive damage to or death of plants due to a separate and uncontrolled pest (or any other cultural factor) may not affect treated and control sets equally; such interactions will frequently lead to an underestimation of effect sizes. A similar result in our demonstration examples would have resulted had the control been based only on noncommercial levels of pesticide application. Objective evaluation of the control may not be possible using the results of individual studies in isolation, and an appeal to historical observations may be required.

As suggested in our introductory comments, one major obstacle to conducting meta-analyses is the availability of appropriate ex-

Fig. 1. Weighted effect sizes and confidence intervals for individual biological control studies and their overall estimate. Means indicated by ●.

Table 8. Comparison of subsets of the biological and chemical control study sets for strawberries.

Control method	Study subset	$1/\hat{\sigma}_{(d)}^2$	d_i ($\sum w_i d_i$)	95% Confidence interval
Biological	All cases	18.116	0.714	0.253, 1.174
	Summer	14.28	0.756	0.246, 1.266
	Winter	3.84	0.555	-0.456, 1.555
	High rate ^z	10.53	0.856	0.252, 1.460
	Low rate	7.58	0.502	-0.210, 1.214
Chemical	All cases	22.270	2.714	2.299, 3.129
	Summer	5.272	1.008	0.167, 1.848
	Winter	17.000	3.241	2.766, 3.716
	High rate ^z	15.523	3.414	2.910, 3.911
	Low rate	6.747	1.083	0.328, 1.838

^zCases with >200,000 predators released per acre (II, III, IV, V, VII, VIII) for biological control and cases with commercial combinations of chemicals (XIII-XVIII).

perimental results in published journal articles. *The minimum information required for quantitative research synthesis includes means, sample sizes, and either standard errors or standard deviations. Likewise, the use of regression analysis or analysis of variance in a meta-analysis requires a complete presentation of mean squares in original articles.* Summaries that provide means and the results of post hoc mean-separation tests or graphs with error bars are inadequate (Wehner and Shaw, 1994). Of course, it is also important that horticulturists report results of studies that produce negative results, as this information will reduce the effects of publication bias.

Literature Cited

Beyer, W.H. (ed.). 1966. Handbook of tables for probability and statistics, p. 280. Chemical Rubber Co., Cleveland.

Cochran, W.G. 1937. Problems arising in the analysis of a series of similar experiments. *J. Royal Statistical Soc.* 4:102-118.

Cochran, W.G. 1938. Some difficulties in the statistical analysis of replicated experiments. *Empire J. Expt. Agr.* 6:157-175.

Fisher, R.A. 1932. Statistical methods for research workers. 4th ed. Oliver and Boyd, London.

Gleser, L.J. and I. Olkin. 1993. Stochastically dependent effect sizes, p. 339-355. In: M. Cooper and L.V. Hedges (eds.). *The handbook of research synthesis.* Russell Sage Foundation, New York.

International Society For Horticultural Science.

1986. Membership summary. Horticultural Research International. Intl. Soc. Hort. Sci., Wageningen, The Netherlands.

*Oatman, E.R., F.E. Gilstrap, R.L. Hale, and V. Voth. 1975. Effect of phorate on the twospotted spider mite, associated predators, and aphids on strawberry in southern California. *Environ. Entomol.* 3:642-644.

*Oatman, E.R., F.E. Gilstrap, and V. Voth. 1976. Effect of different release rates of *Phytoseiulus persimilis* on the two spotted spider mite on strawberry in southern California. *Entomographa* 2:296-273.

*Oatman, E.R. and J.A. McMurtry. 1966. Biological control of two-spotted spider mite in southern California. *J. Econ. Entomol.* 59:433-439.

*Oatman, E.R., J.A. McMurtry, F.E. Gilstrap, and V. Voth. 1977. Effect of releases of *Amblyseius californicus*, *Phytoseiulus persimilis*, and *Typhlodromus occidentalis* on the twospotted spider mite on strawberry in southern California. *J. Econ. Entomol.* 70:45-47.

*Oatman, E.R., J.A. McMurtry, H.H. Shorey, and V. Voth. 1967. Studies on integrating *Phytoseiulus persimilis* releases, chemical applications, cultural manipulations, and natural predation for control of the two-spotted spider mite on strawberry in southern California. *J. Econ. Entomol.* 60:1344-1351.

Olkin, I. and H. Saner. 1993. A robust procedure for combining p-values in integrative research. *Tech. Rpt., Dept. of Statistics, Stanford Univ., Stanford, Calif.*

Pearson, E.S. 1950. A combination of F-statistics. *Technometrics* 37:383-398.

*Pickel, C., N.C. Welch, D. Walsh, and F. Zalom. 1989. Comprehensive research on strawberries,

Annual report: Innovative Lygus and mite control research. Calif. Strawberry Adv. Board Annu. Res. Rpt. p. 122-132.

*Pickel, C., N.C. Welch, and F. Zalom. 1988. Comprehensive research on strawberries, Annual report: Innovative Lygus and mite control. Calif. Strawberry Adv. Board Annu. Res. Rpt. p. 75-95.

Stouffer, S.A., E.A. Buchman, L.C. DeVinney, S.A. Star, and R.M. Williams, Jr. 1949. *The American soldier. vol. I. Adjustment during army life.* Princeton Univ. Press, Princeton, N.J.

Todd, C. 1986. Worldwide evaluation for credibility. *J. Weather Modification* 18(1):49-50.

*Trumble, J.T. and J.P. Morse. 1993. Economics of integrating predacious mites, *Phytoseiulus persimilis*, with pesticides in strawberries. *J. Econ. Entomol.* 86:879-885.

U.S. Department of Agriculture. 1991. *Agricultural statistics. Current Res. Info. Sys.* U.S. Government Printing Office, Washington. D.C.

Wehner, T.C. and D.V. Shaw. 1994. Presentation of analysis of variance results and graphical data. *HortScience* 29:608.

References

Cooper, M. and L.V. Hedges (eds.). 1993. *The handbook of research synthesis.* Russell Sage Foundation, New York.

Hedges, L.V. and I. Olkin. 1985. *Statistical methods for meta analysis.* Academic, Orlando, Fla.

Light, R.J. and D.B. Pillemer. 1984. *Summing up: The science of reviewing research.* Harvard Univ. Press, Cambridge, Mass.