CHAPTER 18

Using Meta-Analyses to Improve the Design of Interventions

Stewart I. Donaldson
Gordon P. Street
Steve Sussman
Nancy S. Tobler

Program development can be greatly enhanced by using systematically collected empirical data to guide design decisions. Empirical data may be used to complement hunches, observations, and anecdotal information gleaned from practice, and these data typically increase the program designer's confidence in the likelihood a program will succeed. This handbook describes in some detail how to use a theory-driven, empirically based approach to develop a program from its initial conception into full-scale program implementation and evaluation. The present chapter will discuss the potential benefits and pitfalls of using data that have been collected from previous research and practice to inform development of a health behavior program. Relevant empirical data from prior work can be used to make the initial conception of the program and selection of program activities as powerful as possible (also see Chapter 7, this volume, on pooling program information).

It may seem like a rather obvious suggestion to use the work of others in program development. That is, standing on the "shoulders of giants" or others who have been down the road one is about to travel forms the basis for most travel guides (this point is made in Chapters 4, 6, and 7, this volume). However, what is
less obvious is how to accurately synthesize and apply this information to the design of a new intervention. A specific focus of this chapter will be placed on using meta-analysis, one of the newest techniques in the family of research synthesis strategies, to guide decisions being made during the program development phase of health behavior research and practice. In part, program development requires gathering and summarizing available data on the efficacy of variations in the contents and delivery modalities relevant to informed program planning. The purpose of meta-analysis is to provide a tight analytic structure to ensure that this process is conducted objectively and to produce results that are easy to interpret.

**META-ANALYSIS**

Research synthesis attempts to illuminate consistencies and document variability in prior studies conducted within a defined program domain (Cooper & Hedges, 1994). Meta-analysis offers a range of statistical techniques that permit pooling findings from a variety of primary studies (i.e., single empirical works such as experiments, quasi-experiments, longitudinal measurement studies, and cross-sectional correlation studies). In contrast to traditional qualitative literature reviews, meta-analysis makes research synthesis an explicit scientific activity in its own right. That is, the meta-analyst goes through a similar set of stages required of any single scientific study: problem formulation, data collection, data evaluation, analysis and interpretation, and public presentation (Cook et al., 1992). Some of the primary characteristics of each of these stages are provided in Table 18.1.

**STRENGTHS OF META-ANALYSIS**

The advantages of using meta-analytic findings to inform program development may be best understood in comparison to the most common alternative sources of empirical information: the single health program study and the traditional qualitative literature review. The major advantage of using meta-analytic findings versus the single health program study is the potential for improved generalizability of findings or external validity (Hall, Rosenhal, Tickle-Degen, & Mosteller, 1994). Although single studies certainly vary on the degree to which findings are likely to generalize beyond their unique characteristics, findings from all primary studies are to some extent bound by the specific participant characteristics, situational factors, and research procedures used. By meta-analyzing findings over a number of primary studies, using different participants in different situations, and using different research procedures, one is able to get
<table>
<thead>
<tr>
<th>Stage Characteristics</th>
<th>Problem Formulation</th>
<th>Data Collection</th>
<th>Data Evaluation</th>
<th>Analysis and Interpretation</th>
<th>Public Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research question asked</td>
<td>What evidence should be included in the review?</td>
<td>What procedures should be used to find relevant evidence?</td>
<td>What retrieved evidence should be included in the review?</td>
<td>What procedures should be used to make inferences about the literature as a whole?</td>
<td>What information should be included in the review report?</td>
</tr>
<tr>
<td>Primary function in review</td>
<td>Constructing definitions that distinguish relevant from irrelevant studies</td>
<td>Determining which sources of potentially relevant studies to examine</td>
<td>Applying criteria to separate “valid” from “invalid” studies</td>
<td>Synthesizing valid retrieved studies</td>
<td>Applying editorial criteria to separate important from unimportant information</td>
</tr>
<tr>
<td>Procedural differences that create variation in review conclusions</td>
<td>1. Differences in included operational definitions</td>
<td>Differences in the research contained in sources of information</td>
<td>1. Differences in quality criteria</td>
<td>Differences in rules of inference</td>
<td>Differences in guidelines for editorial judgment</td>
</tr>
<tr>
<td>2. Differences in operational detail</td>
<td></td>
<td></td>
<td>2. Differences in the influence of nonquality criteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sources of potential invalidity in review conclusions</td>
<td>1. Narrow concepts might make review conclusions less definitive and robust.</td>
<td>1. Accessed studies might be qualitatively different from the target population of studies.</td>
<td>1. Nonquality factors might cause improper weighting of study information.</td>
<td>1. Rules for distinguishing patterns from noise might be inappropriate.</td>
<td>1. Omissions of review procedures might make conclusions irreproducible.</td>
</tr>
<tr>
<td>2. Superficial operational detail might obscure interacting variables.</td>
<td>2. People sampled in accessible studies might be different from target population of people.</td>
<td>2. Omissions in study reports might make conclusions unreliable.</td>
<td>2. Review-based evidence might be used to infer causality.</td>
<td>2. Omission of review findings and study procedures might make conclusions obsolete.</td>
<td></td>
</tr>
</tbody>
</table>

a better estimate of the robustness or the external validity of a given finding or program effect. Thus, in most instances, meta-analytic findings should be a better source of information to use to shape one’s understanding of program theory and to guide aspects of the design of a new intervention than findings from a single, primary study.

Meta-analysis also has important advantages over traditional literature reviews. Most traditional qualitative literature reviews use a “vote count” procedure to accumulate results across studies, which often does not account adequately for (a) statistical power issues (i.e., the ability to detect a true difference or program effect), (b) the methodological rigor of different primary studies, and (c) the possibility that other variables modify basic relationships between a program’s inputs and its outcomes (Cook et al., 1992). For example, the traditional literature reviewer typically classifies how many studies show statistically significant results in the positive direction, negative direction, and how many are not statistically significant. The category with the most votes wins, and conclusions are drawn accordingly. In contrast, the meta-analyst estimates the magnitude of relationships between the intervention and outcomes regardless of whether statistical significance was found, differentially weights studies based on relevant factors such as quality of methodology and sample size, and controls for relevant variables that might influence the bivariate relationship between a program and outcome (e.g., dose, program fidelity). This results in an effect size estimate that can easily be used to determine the practical significance of the program. How to assess practical significance is discussed in more detail in a later section of this chapter on “Using Existing Meta-Analyses.” In summary, meta-analytic reviews typically provide more precise estimates of the relationship between a program and its outcomes than traditional reviews because they overcome statistical conclusion validity problems that are likely to bias program effect estimates.

SOME LIMITATIONS OF META-ANALYSIS

Although meta-analysis has its virtues as a tool for synthesizing complex intervention research literatures, it is not without at least five potentially serious limitations. The main limitation of meta-analysis stems from the problem that it is usually impossible to access all relevant studies or to select a truly random sample of these studies in a particular program domain. For example, meta-analyses of the published literature may overestimate the magnitude of a program effect and possibly may even find a program effect when none exists. This would occur if positive evaluations in a particular area were more likely to be published and null or negative findings were more likely to remain in the investigators’ file drawer, unavailable for research synthesis. In most program domains, it is believed that evaluations that fail to obtain effects often are not published and
hence fail to be considered in research synthesis work, including meta-analysis (Matt & Cook, 1994). If one only synthesizes positive findings, assuming there are a substantial number of null or negative findings in an area, one could imagine finding a positive effect when none would exist if all studies were included. Therefore, it is critical to assess how well a particular meta-analysis has dealt with the “file drawer problem” (Rosenthal, 1995) and related sampling issues before using it to make key program design decisions (see Begg, 1994).

A related sampling problem is that only certain types of quantitative studies get included in meta-analyses. Even among published empirical works, many are excluded from a meta-analysis because the authors of these works fail to include basic statistics necessary for calculation of effect sizes (Sharpe, 1997). In other words, studies that use qualitative methods or that do not use quantitative techniques that easily convert to effect size estimates are systematically excluded from meta-analysis, thus limiting the generalizability of meta-analytic findings. Instead of simply excluding such studies from meta-analyses, corrective procedures and calculations have been used to produce an effect size estimator for a study. Ray and Shadish (1996) have demonstrated that some of these calculations produce overestimates of effect sizes, whereas others produce underestimates. Consequently, the methods of producing effect sizes for a meta-analysis also should be evaluated.

Another issue regarding different studies that are included in meta-analysis pertains to the study context dependence of effect size estimates. All else being equal (including the size of the true treatment effect), two studies that use different comparison groups will produce different effect size estimates. For example, some meta-analyses will combine studies that compare a treatment against no treatment (e.g., wait list control), but others will compare a treatment against an alternative treatment. The control group may show naturally occurring change or no change. However, the alternative treatment often will result in some change from baseline, which may result in a smaller effect size estimate. This is particularly important when evaluating meta-analyses that include single-group, pre-treatment-posttreatment studies because they have no real comparison group at all. Instead, the pretreatment mean and standard deviation often are used as an estimate of a comparison group. Lipsey and Wilson (1993) found that effect sizes produced from single-group studies were 61% larger than those studies using control or comparison groups. Meta-analyses may or may not analyze and report separately these three types of comparisons to calculate an effect size (i.e., treatment vs. comparison, treatment vs. control, posttreatment vs. pretreatment).

Two other common threats to the validity of meta-analytic findings are problems due to mixing dissimilar studies and the inclusion of poor-quality studies (Sharpe, 1997). The first of these threats has been labeled mixing apples and oranges. Meta-analyses that combine statistical results from studies that manipulate and measure different variables among different participant populations
may end up with a meaningless superordinate analysis. The key to overcoming this problem is to make sure the meta-analyst has selected a reasonable collection of studies to include in the pool—striking a balance between too broad and too narrow (Wartman, 1994). The second threat has been called “garbage in, garbage out” (Eysenck, 1978). In an effort to overcome study selection bias, some meta-analysis gather all available studies (e.g., Smith & Glass, 1977). Critics argue that including methodologically weak studies provides a distorted picture of the issue under investigation. It is important to consider how well a particular meta-analysis addresses the issue of variability in study quality. One other caveat: We cannot with any confidence ascribe causality to a relationship based on meta-analysis alone. The boundaries to how much we can learn about causation with meta-analysis are defined by the underlying primary studies (Wartman, 1994). Despite these potential limitations, meta-analyses that contain a thoughtful combination of qualitative and quantitative insights can dramatically improve the initial conception of a program, which can then be refined using an empirical program development approach (see Chapter 19, this volume; Sussman, Petosa, & Clark, 1996).

META-ANALYSIS AND PROGRAM DEVELOPMENT

HOW TO INTERPRET THE RESULTS OF EXISTING META-ANALYSES

Calculations. This section describes the specifics of how to use meta-analysis for improving program development. In its best form, meta-analysis can provide program designers an accurate estimate of the relationships between a specific type of program and its desired outcomes across a range of implementation environments and participants. This estimate typically is expressed as an estimate of the effect size of the program on the outcome. The effect size is a population parameter that cannot be determined directly. Rather, it is an inference to a population based on sample distributions. It is estimated from the sample data collected as part of the primary studies. Statistical summaries of primary study data are used as single data points in a meta-analysis. Effect size is the ratio of the size of the difference between treatment (M1) and control group (M0) means over the common standard deviation (see Figure 18.1). Simply stated, the effect size estimate derived from a typical meta-analysis is the average effect size estimate across a range of primary studies.

It is important to evaluate the overall mean effect size of a treatment for variability and for appropriateness to the program that one hopes to develop. The standard deviation of the mean effect size and the range of the individual effect
sizes are typically presented in a meta-analysis. Also, the individual effect sizes and confidence intervals occasionally are presented for each study included in the meta-analysis. These data summaries can provide a rough indication of the range of outcomes that one might expect from a planned program. Sometimes, this variability may be considerable and raise concern that, despite a respectable overall effect size, a treatment may produce little to no benefit at all in certain contexts. One may want to examine the descriptions of the individual studies to determine which ones are more representative of the parameter that one's own program requires (studies that have similar populations, clinicians, etc.). Then one can examine the range and confidence intervals of these studies and calculate the mean effect size and standard deviation for this more focused subset of studies that are more relevant to one's program.

Once a program designer finds meta-analyses relevant to the new program being planned, she or he must be able to interpret the practical significance of these findings. Lipsey (1990) provides a framework for understanding the practical significance of effect size estimates of program effects. On the basis of the
cumulative distribution of 102 selected mean effect sizes from 186 meta-analyses of program effectiveness studies (a meta-analysis of meta-analyses), Lipsey shows that an effect size between .00 and .32 should be considered small, between .33 and .55 should be considered medium, and between .56 to 1.20 should be considered a large effect in program effectiveness research. The wide, large effect category illustrates that some programs can have very strong effects. It is important to point out that this is a general framework for interpreting effect size estimates across a wide variety of program areas. The range for each category may vary somewhat based on the specific program domain. Nevertheless, this framework is useful to determine whether a specific type of program has demonstrated practically significant findings in previous evaluations.

One of the virtues of effect size estimates derived from meta-analyses is that they are easily converted into other common statistics, which is very useful for practical interpretation purposes. Table 18.2 provides effect size equivalents for common indices of strength of association ($r^2$ and $r$), Cohen’s (1977) U3 measure, and Rosenthal and Rubin’s (1982) “binomial effect size display” (BESD). Each of these statistics can be used to help the program developer understand the practical significance of findings from prior program effectiveness research in the program domain of interest.

For example, if a program developer locates a meta-analysis that estimates the effect size of a program similar to one planned to be .50 (a medium effect, as described earlier), she or he can use Table 18.2 to determine that this finding estimates that 6% of the variance ($r^2 = .06$) in the outcome is accounted for by the program. The planner can expect that his or her program will account for 6% of the variation in the health behavior. The next column shows that this statistic is equivalent to finding a correlation coefficient of $r = .24$ between the program and the outcome. Furthermore, Cohen’s U3 measure (same row, third column in Table 18.2) estimates that 69% of the program or treatment group scored higher than the control group mean on the outcome measure of interest. (By comparison, if there was no effect [ES = .00], only 50% of the program group would have scored higher than the control group mean, and if ES = 1.2, 88% of the program group would have scored higher than the control group mean.)

A variation on Cohen’s U3 percentage overlap index is the BESD (Rosenthal & Rubin, 1982). Rosenthal and Rubin suggest that for the purposes of illustrating ES, the success threshold can be presumed to be at the grand median for the combined treatment and control group distributions. This grand median standard presumes that when there is no effect (ES = .00), there will be a 50-50 success-failure split and a widening difference as the ES increases. In the example above, for an ES = .50, 62% of the program group would be considered successful, whereas only 38% of the control group would be successful (see column 4 in Table 18.2). Furthermore, this corresponds to a 24% differential between the two groups, as is shown in column 5. Consider a practical, concrete example for col-
<table>
<thead>
<tr>
<th>ES</th>
<th>PV (r')</th>
<th>r</th>
<th>U3: % of T Above X</th>
<th>BESD C vs. T Success Rates</th>
<th>BESD C vs. T Differential</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.002</td>
<td>.05</td>
<td>54</td>
<td>.47</td>
<td>.52</td>
</tr>
<tr>
<td>0.2</td>
<td>.01</td>
<td>.10</td>
<td>58</td>
<td>.45</td>
<td>.55</td>
</tr>
<tr>
<td>0.3</td>
<td>.02</td>
<td>.15</td>
<td>62</td>
<td>.42</td>
<td>.57</td>
</tr>
<tr>
<td>0.4</td>
<td>.04</td>
<td>.20</td>
<td>66</td>
<td>.40</td>
<td>.60</td>
</tr>
<tr>
<td>0.5</td>
<td>.06</td>
<td>.24</td>
<td>69</td>
<td>.38</td>
<td>.62</td>
</tr>
<tr>
<td>0.6</td>
<td>.08</td>
<td>.29</td>
<td>73</td>
<td>.35</td>
<td>.64</td>
</tr>
<tr>
<td>0.7</td>
<td>.11</td>
<td>.33</td>
<td>76</td>
<td>.33</td>
<td>.66</td>
</tr>
<tr>
<td>0.8</td>
<td>.14</td>
<td>.37</td>
<td>79</td>
<td>.31</td>
<td>.68</td>
</tr>
<tr>
<td>0.9</td>
<td>.17</td>
<td>.41</td>
<td>82</td>
<td>.29</td>
<td>.70</td>
</tr>
<tr>
<td>1.0</td>
<td>.20</td>
<td>.45</td>
<td>84</td>
<td>.27</td>
<td>.72</td>
</tr>
<tr>
<td>1.1</td>
<td>.23</td>
<td>.48</td>
<td>86</td>
<td>.26</td>
<td>.74</td>
</tr>
<tr>
<td>1.2</td>
<td>.26</td>
<td>.51</td>
<td>88</td>
<td>.24</td>
<td>.75</td>
</tr>
<tr>
<td>1.3</td>
<td>.30</td>
<td>.54</td>
<td>90</td>
<td>.23</td>
<td>.77</td>
</tr>
<tr>
<td>1.4</td>
<td>.33</td>
<td>.57</td>
<td>92</td>
<td>.21</td>
<td>.78</td>
</tr>
<tr>
<td>1.5</td>
<td>.36</td>
<td>.60</td>
<td>93</td>
<td>.20</td>
<td>.80</td>
</tr>
<tr>
<td>1.6</td>
<td>.39</td>
<td>.62</td>
<td>95</td>
<td>.19</td>
<td>.81</td>
</tr>
<tr>
<td>1.7</td>
<td>.42</td>
<td>.65</td>
<td>96</td>
<td>.17</td>
<td>.82</td>
</tr>
<tr>
<td>1.8</td>
<td>.45</td>
<td>.67</td>
<td>96</td>
<td>.16</td>
<td>.83</td>
</tr>
<tr>
<td>1.9</td>
<td>.47</td>
<td>.69</td>
<td>97</td>
<td>.15</td>
<td>.84</td>
</tr>
<tr>
<td>2.0</td>
<td>.50</td>
<td>.71</td>
<td>98</td>
<td>.14</td>
<td>.85</td>
</tr>
<tr>
<td>2.1</td>
<td>.52</td>
<td>.72</td>
<td>98</td>
<td>.14</td>
<td>.86</td>
</tr>
<tr>
<td>2.2</td>
<td>.55</td>
<td>.74</td>
<td>99</td>
<td>.13</td>
<td>.87</td>
</tr>
<tr>
<td>2.3</td>
<td>.57</td>
<td>.75</td>
<td>99</td>
<td>.12</td>
<td>.87</td>
</tr>
<tr>
<td>2.4</td>
<td>.59</td>
<td>.77</td>
<td>99</td>
<td>.11</td>
<td>.88</td>
</tr>
<tr>
<td>2.5</td>
<td>.61</td>
<td>.78</td>
<td>99</td>
<td>.11</td>
<td>.89</td>
</tr>
<tr>
<td>2.6</td>
<td>.63</td>
<td>.79</td>
<td>99</td>
<td>.10</td>
<td>.89</td>
</tr>
<tr>
<td>2.7</td>
<td>.65</td>
<td>.80</td>
<td>99</td>
<td>.10</td>
<td>.90</td>
</tr>
<tr>
<td>2.8</td>
<td>.66</td>
<td>.81</td>
<td>99</td>
<td>.09</td>
<td>.90</td>
</tr>
<tr>
<td>2.9</td>
<td>.68</td>
<td>.82</td>
<td>99</td>
<td>.09</td>
<td>.91</td>
</tr>
<tr>
<td>3.0</td>
<td>.69</td>
<td>.83</td>
<td>99</td>
<td>.08</td>
<td>.91</td>
</tr>
</tbody>
</table>
columns 4 and 5. If one were talking about a smoking cessation program \((N = 200)\) smokers in the program compared to \(N = 200\) smokers in the control or comparison group), this would mean that 124 smokers who participated in the program of interest stopped smoking, in contrast to only 76 in the comparison program. In this context, a medium effect size of .50 appears “practically” significant. It shows that 48 more people stopped smoking in the program group compared to the control group (i.e., the success differential shown in column 5 in Table 18.2).

Of course, presentation of these different estimates of effect sizes in the absence of other information can be misleading. A small effect may have enormous implications for health if it can save many lives. Because meta-analyses may overestimate effects of programming (null effects being less likely to appear in the literature), this problem is less likely to result from meta-analytic interpretation. However, if one simply looks at effect size as a rule for selection, other important information can get lost. For example, disseminability of a program is important to consider in addition to effect size. Certainly, effect size measures will continue to be a focus of discussion among statisticians and program developers. The measures provided herein are quite useful if one keeps in mind these other issues.

Using These Calculations to Plan New Programs. Having shown how to interpret the practical significance of effect size estimates, we will now consider the kinds of relationships a program designer might examine during program development. These relations provide a means to select feasible program components for new contexts. Four potential relations are identified here: strength of program effects over a range of program-constituent information, program-mediator links, mediator-outcome links, and moderator effects. The most common information available is the relationship between a specific type of program and its desired outcomes. The two most common aspects of a program that may be examined or varied are program content (e.g., consequences oriented vs. acquisition oriented) and type of delivery (e.g., didactic vs. interactive). Meta-analyses can be very useful for showing that over a range of implementation type and fidelity, specific content packages produce significantly stronger program effects. Of course, when these kinds of patterns emerge, they can be used to select or eliminate potential content areas and delivery approaches for the new program being planned (see Tobler & Stratton, 1997).

Many programs are conceptualized to consist of a number of components, often intended to lead to immediate or proximal changes (mediators) in participants that are presumed to later cause more distal intended outcomes (see Figure 18.2). Hansen and McNeal (1996) suggested that health behavior programs attempt to change these immediate factors, which, in turn, elicit desired outcomes. The outcomes typically studied are indirectly influenced by the program and are the result of whether the mediating factors are changed by the program and are
linked to the desired outcome. They call this the \textit{law of indirect effect} in health behavior programming (see Chapter 19, this volume). Of course, some program, mediator, and outcome links occur quite proximally in time (e.g., smoking cessation), whereas other such links show long delays between changes in the mediator and observable changes in behavior (e.g., smoking prevention). Meta-analyses also can be useful for estimating the effect size between a program and its proximal outcomes (mediators) presumed to lead to desired health behavior change. Program-mediator “strength-of-link” information is very important to the program developer who is searching for program components that might alter hypothesized mediators of change in a certain context.

Meta-analyses also may help determine whether the link between the mediators (as shown in Figure 18.2) and the outcomes is likely to occur in a new program application. These links are critical because they ultimately determine how likely it is that a successful program, through changing mediator variables, is likely to produce health behavior change. Hansen and McNeal (1996) described this as the \textit{law of maximum expected potential effect}, which suggests that the magnitude of change in a behavioral outcome that a program can produce is directly limited by the strength of relationships that exist between mediators and tar-
geted behaviors (see Chapter 19, this volume). Meta-analyses of correlations or other measures of association between presumed mediators and outcomes, as opposed to the standardized difference effect sizes discussed in detail earlier, can be useful for estimating the maximum potential of a program to affect desired outcomes. Furthermore, Donaldson (Chapter 19, this volume) illustrates how moderator variables also are critical to an overall understanding of program development. Meta-analyses also may be used to estimate potential moderator effects in the new program. In particular, this information can be used to consider whether a program is likely to affect different genders or other subgroups within a population.

CONDUCTING ONE'S OWN META-ANALYSIS OF PREVIOUS RESEARCH

In some cases, previously published meta-analyses may have done much of the work if they overlap or contain a subset of studies that address one's area of interest. However, if one cannot find any meta-analytic reviews that adequately address the intervention program elements that one is attempting to assess, one may benefit from conducting a meta-analysis of one's own, if only a cursory one. Depending on the program development goals, this task may not be as difficult or time-consuming as it might sound, particularly if the focus is narrow. The procedures are fairly simple, and summaries of varying length and complexity have been published that can provide guidance (e.g., Durlak & Lipsey, 1991; Glass, McGaw, & Smith, 1981; Rosenthal, 1995).

Conducting a meta-analysis is much like conducting an individual study or evaluation of a program, except that the data are the effect sizes, and each subject is a study or evaluation rather than groups or individuals. Durlak and Lipsey (1991) conceptualize meta-analysis as composed of six major steps: (a) formulating the question, (b) searching for studies, (c) coding variables, (d) calculating effect sizes, (e) analyzing the data, and (f) drawing conclusions and interpretations. Tracking down all of the relevant studies for a meta-analysis for publication can take considerable time. In program development, one's time may be relatively limited, so one's sample of studies may be more skewed by ease of availability. Therefore, one should start as soon as possible to gather every study one can find relevant to the question, including unpublished studies. Excluding studies that failed to get published because they did not find significant results can result in a meta-analysis that overestimates the effectiveness of an intervention (Smith, 1980; as discussed previously herein). One should not rely only on computer database searches because researchers have found them unreliable, tending to miss as many as two out of three relevant studies (Durlak & Lipsey, 1991). One should supplement computer database searches by reviewing reference lists of relevant
studies that one has already found and by contacting experts in the field for their published and unpublished studies and recommendations of additional studies to review.

Perhaps the most difficult and time-consuming step is coding the variables that will serve as the predictors and covariates in the meta-analysis. The more variables one determines are necessary to code, the longer this process can take, often averaging hours for each study in a published meta-analysis, so one should choose carefully. Variables that can be coded include demographic characteristics of participants (e.g., age, sex, ethnicity, education level, cognitive ability), intervention characteristics (e.g., duration, setting, mode of administration), and methodological considerations (e.g., research design, type of control group).

Because most intervention research compares data between groups, intervention study meta-analyses typically use a standardized group difference effect size statistic. Calculating this statistic actually can be quite easy if the studies provide three basic statistics for each group in the study: means, standard deviations, and sample sizes. The statistic is basically the difference between the means of two groups (e.g., a control or comparison group and a treatment group) divided by their pooled standard deviations. There are different ways to calculate the pooled standard deviations, depending on theoretical issues, but the most commonly used calculation is to multiply the standard deviation of each group by one less than its sample size, add the result, and then divide that by two less than the sum of both samples.

However, some studies do not report the necessary statistics. Alternative methods of calculating or estimating effect size must be used with recognition that the resulting effect sizes may not actually be equivalent to those produced using the basic formula (Ray & Shadish, 1996). If one is drawing studies from previously published meta-analyses that used standardized group mean effect sizes, one can borrow their effect size statistics rather than recalculating them with the recognition that the same cautions apply. Also, many studies may not use the same type of comparison group or sometimes any comparison group at all. In these cases, it will be important to include type of comparison group in the coding of variables so that differences between them can be analyzed and evaluated (i.e., separate analysis).

One analyzes effect sizes just like the dependent variables in individual studies. First, basic statistics (e.g., means, standard deviations, standard errors) should be calculated for the effect sizes across all studies and within important subsets of the studies. Rosenthal (1995) suggests calculating confidence intervals around each of the resulting average effect sizes to provide a sense of how big or small the overall effect of a treatment might really be across the studies or across subsets. Using a simple statistical calculator, one can enter the effect sizes of each study for the overall sample or subsets of the sample, and the mean and standard error can be produced. The 95% confidence interval is equal to the mean plus or
minus two times the standard error. These confidence intervals allow quick evaluations regarding whether an effect size will be found significantly different from either another effect size or zero in further analyses.

Then, statistical analyses of the differences between those subsets can be conducted. For example, if differences in effectiveness between genders are a concern, an ANOVA could be conducted to determine if the effect sizes differed between studies that examined only women, those that examined only men, and those that included both men and women. T tests or contrasts could be used to assess differences between pairs or combinations of those subsets. Similarly, the same statistics could be run within studies that use different control groups (e.g., wait list control, pretest-posttest design) or other potentially influential methodological differences. Again, using the procedures described earlier can be very helpful when one cannot find any meta-analytic reviews that adequately address one's specific information needs.

USE OF META-ANALYSIS AS A MEANS OF EVALUATING COMPONENT OR PILOT STUDIES

Meta-analysis also presents program developers with new approaches for evaluating the success of a series of program components or pilot tests (these types of studies were discussed in Chapters 13 and 16, this volume). Such studies often are tested in the absence of a control group. A very simple approach would be to compare outcomes of program components or a pilot test directly to the BESD treatment group success rate estimates (see Table 18.2) that would be expected based on mean effect size for a treatment obtained from meta-analysis. As noted earlier, a treatment that is shown in meta-analysis to have ES = .50 would be expected to produce successful outcomes for 62% of those who received the treatment; a similar impact could be expected from a program component or pilot test. More realistically, variability due to random variation could be accounted for by determining ranges of expected outcomes. For a program component or pilot test that achieves an ES of .50 on a mediator variable and an ES standard deviation of .10, a successful outcome from 60% to 64% of program component or pilot test recipients might be expected to occur two thirds of the time. If the impact of the program component or pilot test has been determined, an estimated effect size can be determined for it from Table 18.2 and evaluated to determine if the obtained effect is falling within the expected effect size range. These comparisons would be quite rough, of course.

Another approach would be to conduct rudimentary meta-analysis on the component or series of components or the pilot test itself, using individual implementations (i.e., sequential trials) or components of the implementation (e.g., different sites, individual clinicians) to represent separate component or pilot
Using Meta-Analyses

studies. Pretreatment to posttreatment immediate-impact effects could be evaluated within each of these implementations, and effect sizes and confidence intervals could be calculated based on the differences between them. The purpose of evaluation is to determine how big or small the effects are for each component or pilot study implementation, how much variability there is between them, and how much consistency there is between them. If the implementations consistently produce positive effect sizes (subjects are changing in the desired direction on mediator variables) and the confidence intervals overlap above zero, the evidence of this meta-analysis suggests that the program component or pilot test version of a complete program is having some beneficial effect. If there are substantial differences between the effect sizes, the program developer can determine what is different between these implementations, determine which features promote better outcomes for clients, and modify existing or future implementations.

**SOME EXAMPLES OF META-ANALYSIS IN HEALTH BEHAVIOR RESEARCH AND PRACTICE**

Over the past two decades, conventional qualitative analyses of the literature have not yielded convincing evidence for the efficacy of psychological, educational, and behavioral programs, collectively known as social programs. In fact, Rossi and Wright (1984) described the field of evaluation research as a “parade of close-to-zero effects.” Rossi (1985) sardonically coined a series of laws that reflect the difficulties of producing social change through programs. The metallic laws of evaluation include the following:

1. **The iron law**: The expected value of any net impact assessment of any social program is zero.

2. **The stainless steel law**: The better designed the impact assessment of a social program, the more likely is the net impact to be zero.

3. **The copper law**: The more social programs are designed to change individuals, the more likely the net impact will be zero.

However, Lipsey and Wilson (1993) analyzed essentially the same evaluation literature using meta-analytic techniques and came to a very different conclusion about the field of program development. Meta-analytic reviews show a strong, dramatic pattern of positive overall effects that cannot readily be explained as artifacts of meta-analytic technique or generalized placebo effects.
More specifically, the most rigorous assessment of 156 meta-analyses, encompassing approximately 9,400 individual program effectiveness studies and more than 1 million individual participants, shows the average program effect size to be .47. Included in this work are a large number of health behavior meta-analyses, including, for example, the following:

1. Patient education about treatment regimens, preventive behavior, self-care, and related topics—mean ES = .74 for all outcomes (Posavac, 1980);
2. Physician-delivered smoking cessation/reduction programs—mean ES = .34 for quit rates (Dotson, 1990);
3. Worksite smoking cessation/reduction programs—mean ES = .34 for quit rates (Fisher, 1990);
4. Nonmedical psychologically based treatment of chronic pain—mean ES = 1.10 (Malone, Strube, & Scogin, 1989);
5. Multidisciplinary treatments for chronic back pain—mean ES = 1.25 all outcomes (Flor, Fydrich, & Turk, 1992);
6. Exercise interventions for depression—mean ES = .54 on depression (North, 1989);

This research by Lipsey and Wilson (1993) is one of the most powerful examples of the value of meta-analysis for informing program development.

Other useful examples of the value of meta-analysis in health behavior programming can be found in the school-based drug abuse prevention area (see Tobler & Stratton, 1997). Meta-analysis in this arena has demonstrated knowledge and attitude changes for alcohol, tobacco, marijuana, and other illicit drug use. More important, although the magnitude for drug use behaviors is less, certain types of interventions are differentially effective in preventing, delaying, or decreasing the use of drugs. Combining short- and long-term effects across smoking and alcohol programs, Rundall and Bruvold (1988) showed a mean effect size for smoking behaviors of .25 (n = 41) and .14 (n = 19) for alcohol behaviors. Bangert-Drowns (1988) found a mean effect of .12 for 14 programs that included drug use measures. Tobler et al. (in press) meta-analyzed 207 programs and found that interactive programs that emphasize interactions and exchange with peers were statistically superior to noninteractive programs such as Drug Abuse Resistance Education (DARE) for all adolescents, including minority populations. In a subset of 93 high-quality, well-implemented programs, the interac-
tive programs’ mean effect size was .16 compared to .03 for the noninteractive programs. Using Rosenthal and Rubin’s (1982) binomial effect size display, these modest effect sizes equal success rates of 8% and 1%, respectively. These meta-analytic findings suggest the value of designing future adolescent drug prevention programs to contain an interactive format.

CONCLUSION

Most would agree that developing effective health behavior programs is a challenging endeavor. It has been argued in this chapter that program development can be greatly enhanced by using meta-analyses of prior empirical studies to help guide program design decisions. Findings from systematic research syntheses should be used to complement, rather than replace, more qualitative descriptions and understanding of the new program content, context, participants, program delivery staff, and the like. This chapter has attempted to provide a balanced account of the strengths and weaknesses of using meta-analysis in program development. Furthermore, how to use existing meta-analyses and conduct one’s own meta-analysis for program development purposes was described in some detail. Examples of meta-analyses from the health behavior literature were provided to illustrate many of the key points. It is hoped that the issues discussed in this chapter will be helpful toward the effort to develop more effective programs based on behavior science knowledge and principles, for the purpose of preventing and ameliorating some of our most pressing health and social concerns.

REFERENCES


I have long been my belief that the most common reason for the failure of health behavior change programs has been poor choices of intervention strategies. Furthermore, I believe that these poor choices could have been avoided had the program planners done their homework in reading the available literature in the field. All too often, we seem not merely to persist in reinventing the wheel but to persevere in adopting the square model for our wheel in ignorance of others' past failures with square wheels. The general failure of planners and policymakers to use the available research has been discussed by many authors (Galster, 1999; Getting Smart, 1995; Hanson, 1997).

In teaching program planning, I have emphasized that a thorough literature review was an essential early step. I have found, however, that it is not enough simply to read the literature. It is necessary for planners to make a judgment based on that literature about what type of intervention is most likely to achieve the goals of their program with the population and setting they hope to serve.

Unfortunately, many planners seem to have little idea of how to make that judgment (Getting Smart, 1995). Some approach it as if they were choosing lunch in a cafeteria. One intervention looks particularly appetizing to them so they choose it. They do not examine the evidence for its effectiveness or seek any comparison between it and other alternatives. The literature, for them, is merely an array of choices from which they choose based on personal attraction.

Others behave like the carpenter whose only tool is a chisel. Just as the carpenter struggles to pound nails and cut lumber with his chisel, these persons try to achieve every goal with the same programmatic strategy. One may approach every problem armed with public service announcements, billboards, posters, and other mass media. Another tries to resolve every health problem with group training in stress management and self-esteem promotion. They search the literature only to find examples of the use of their favorite approach, not to compare the effectiveness of that approach to others.

Donaldson, Street, Sussman, and Tobler present meta-analysis as a systematic guide to making those critical choices among intervention strategies. I fully agree that meta-analysis can greatly enhance this process. There are, however, several additional points I think the reader should be aware of regarding the use of meta-analysis in selecting interventions. First, I want to revisit the issue of practical significance of meta-analysis findings. Most readers should be aware that a finding of statistical significance is quite a different thing from finding that the result is important. Significance only reflects our degree of confidence that there truly is a difference between two or more groups on some measure; it does not mean that the difference is great enough to be of any practical importance. In a large sample, one group of men may be found to be significantly taller than another group of men, but the mean difference in their heights could be as little as one inch.
Donaldson et al. describe effect sizes between .00 and .32 as small, between .33 and .55 as medium, and between .56 and 1.20 as large. An effect size greater than .55 certainly is likely to be statistically significant and very often also will have practical significance but not in all cases. In situations in which the preintervention variance was small, a change of one standard deviation may not be of much practical effect. Donaldson et al., for instance, note that an effect size of .50 represents only an explanation of 6% of the variance in outcomes. Although this may be perfectly adequate in some cases, 6% may be far too little bang for the buck in other cases.

This is particularly true when the measure in use suffers from a plateau effect in the population or when the measure used has only a narrow range of possible scores. Take, for instance, an outcome measure of belief in the dangerousness of cocaine, which used a 6-point scale with points identified as very safe, safe, somewhat safe, somewhat dangerous, dangerous, and very dangerous. The preintervention mean might be 5.65 with a standard deviation of 0.40. An increase in the mean danger rating of 0.30 would be a large effect size and probably would be significant, but it would be only a change from an average rating of very dangerous to an average rating still of very dangerous (just a slightly more dangerous perception of cocaine).

Second, a rather more important problem is the likelihood that effect sizes found by meta-analysis will prove to be overestimates of the effect of the intervention in a real program. In the health care field, for instance, although Chalmers et al. (1987) report that meta-analytic conclusions about the effectiveness of treatments have been consistently supported by evidence from large clinical trials, the impact of these treatments on the health of the population has been far less than predicted. The General Accounting Office (Silberman, Droitcour, & Scullin, 1992) has documented the consistent failure of new medical treatments to reduce death rates in the population to the extent predicted from the research data.

There are several possible explanations for these discrepancies between meta-analytic results and real-world experience. Subjects included in clinical trials and other studies often will have been carefully screened for suitability, resulting in the intervention being administered only to persons most likely to benefit from it. Being aware of the danger of Type III error (Basch, Slepcevich, Gold, Duncan, & Kolbe, 1985; Steckler, 1989), the researchers may have taken careful precautions to guarantee that the intervention was fully implemented and that it was implemented by highly qualified personnel—conditions that may not be matched in the real world. It is also possible that the intensity of the intervention may have been greater in research settings than it will be in later community applications. Furthermore, studies in which the intervention failed to produce the expected results are less likely to have been published—the so-called file cabinet effect.

Finally, it is important for the planner to look at more than just the intervention and the outcomes studied in a meta-analysis. It is also important to examine the target populations and the settings in which the research studies were conducted and the impacts, if any, that these had on the effect sizes. It is important, for instance, that one should not select an intervention that has been found to be highly effective only for in-school populations of good readers if one's program will be community based and focused on academically unsuccessful adolescents.

Study design characteristics also should be considered in any meta-evaluation. These include all of the systematic aspects of the research design and procedure, except those that are part of the study context. Do not put your faith in an intervention that produced large effect sizes in only the weakest research designs but only small effect sizes in well-controlled experiments. Meta-analysis is a valuable tool for synthesizing the results of many studies in a way that can help us to make program-planning decisions. We need to be aware of the limitations of this approach to make intelligent use of it. With due regard for these limits, it will help us plan better interventions.


