Statistical methodologies to pool across multiple intervention studies

Authors

- Shrikant I. Bangdiwala, Department of Biostatistics, University of North Carolina at Chapel Hill, NC, USA, & Institute for Social and Health Sciences, University of South Africa, Johannesburg, South Africa
- Alok Bhargava, University of Maryland School of Public Policy, College Park, MD, USA
- Daniel P. O'Connor, Texas Obesity Research Center, Department of Health and Human Performance, University of Houston, Houston, TX, USA
- Thomas N. Robinson, Center for Healthy Weight, Stanford University, Stanford, CA, USA
- Susan Michie, Centre for Behaviour Change & Department of Clinical, Educational and Health Psychology, University College London, UK
- David M. Murray, Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD, USA
- June Stevens, Department of Nutrition, University of North Carolina at Chapel Hill, NC, USA
- Steven H. Belle, Department of Epidemiology, University of Pittsburgh, PA, USA
- Thomas N. Templin, Office of Health Research, , College of Nursing, Wayne State University, Detroit, MI, USA
- Charlotte A. Pratt, Division of Cardiovascular Sciences, National Heart Lung and Blood Institute, Bethesda, MD, USA

Word count: 4433

Introduction

Randomized controlled trials (RCT) are considered the gold standard experimental study design for establishing the causal effect of an intervention on an outcome of interest. RCTs are usually designed to have high internal validity in addressing specific hypotheses but may have less external validity as their inclusion and exclusion criteria may be very restrictive. Often there are many similar trials addressing the same type of research hypotheses, but with different target populations, settings or outcome measures. Such trials may not evaluate exactly the same intervention, especially in trials of interventions that include combinations of multiple behavioral, social, pharmacological and/or environmental components.

A question to consider is whether there are benefits from combining data from several studies. The combining of data from various randomized controlled trials (RCTs) can be useful in applications beyond estimation of the overall intervention effect. For example, it may be informative to combine the data for increasing sample sizes of subgroups in which to examine the intervention effect, or to increase the number of events for secondary outcomes, or to reduce variances and obtain more precise confidence intervals for outcomes and adverse events. An alternative to combining the results from various small trials would be to undertake a large definitive trial, i.e., one that establishes conclusively the safety and efficacy of a proposed intervention. However, such trials are not always feasible due to requiring very large sample sizes, long duration, large costs, or by the nature of the intervention (e.g. policy interventions).

In many situations, RCTs are of multi-component interventions aimed at preventing conditions such as diabetes and obesity, or for subjects having a high cardiovascular risk profile. Combining and analyzing the data from heterogeneous randomized controlled trials of complex multiple-component intervention studies, or discussing them in a systematic review, is not straightforward. The first important question is whether it is appropriate at all to combine data from a set of heterogeneous randomized controlled trials. Once the decision to combine the data or results of the various trials is made, the issue of how to combine the trials needs to be considered. A review of possible procedures concluded that the most serious methodological limitation is the question of what studies should be combined rather than how to combine them (DeMets, 1987).

The objective of this manuscript is to describe certain statistical issues to be considered when combining data across studies, especially studies that share many commonalities, as in consortia studies. The present article describes certain aspects to be considered when combining data across studies, that were discussed in an NIH sponsored workshop on 'pooling issues across studies in consortia' (see Belle et al in this issue). Several statistical methodologies are described and their advantages and limitations are explored. In addition, illustrations of combing data are given with reference to examples from the Childhood Obesity Prevention and Treatment Research (COPTR) consortium.

Methodological issues in combining similar but different intervention studies

To be combined, trials should address the same, or similar, research question(s) in similar populations and settings using similar intervention components and implementation approaches, and having the same or similar outcome variables. However, strict inclusion criteria that attempt to define trials that are 'very similar' may lead to an overly conservative decision that trials should not be combined unless all components are identical in all studies (Spinks et al 2009). The more aspects they share in common, such as conceptual or theoretical framework, inclusion and exclusion criteria, recruitment methods, measures, timing of assessments, intervention approaches, procedures of study implementation (e.g. training, quality assurance, and data management), the less heterogeneity and the more convincing the argument for combining will be. For example, tight definitions of behavioral therapies, or classifying the components using a common taxonomy, may better define exposure variables and strengthen the argument for combining across behavioral interventions.

An example of addressing this question comes from consortia funded by the National Institutes of Health to test the efficacy of a diverse set of obesity-related interventions at multiple sites across the country. The consortia include the Childhood Obesity Prevention and Treatment Research (COPTR; 4 studies) (Pratt et al 2013), the Early Adult Reduction of weight through LifestYle intervention (EARLY; 7 studies) (Lytle et al 2014), the Obesity Related Behavioral Intervention Trials (ORBIT; 7 studies) (Czajkowski et al 2015), and the Lifestyle Interventions For Expectant Moms (LIFE-Moms, 7 studies). COPTR is testing multi-level intervention approaches to prevent excess weight gain in youth, and to reduce weight among overweight and obese youth. Targeted age

groups are preschoolers (2-5 year olds), pre-adolescents and adolescents (7-14 year olds) of diverse racial and ethnic groups in four different locations in the U.S. EARLY is testing innovative behavioral approaches for weight control in young adults, 18-35 years of age, at high risk for weight gain. ORBIT is testing methods to translate findings from basic research on human behavior into more effective clinical, community, and population interventions to reduce obesity in a diverse group of subjects. LIFE-Moms is testing behavioral/lifestyle interventions in overweight and obese pregnant women designed to improve weight and metabolic outcomes among women and their children.

Within each of these consortia, the individual trials are each designed to be stand-alone studies with adequate power to address their respective primary hypotheses. There is interest to combine study data for several reasons: the potential to explore certain important secondary hypotheses that are not testable in any one study (i.e., new research questions, such as testing for geographical and other contextual effects that are typically constant within a single study), and more can potentially be learned from information across trials than the information available from each individual study. For example, it may allow the investigation of effect modification by type of study approach or by population or contextual characteristics. Because the studies within consortia have different study populations and intervention approaches, there are analytic challenges in exploring relationships when combining studies, even though they may all have a common outcome measure.

There are both potential advantages and disadvantages of combining data across studies. Potential advantages are larger sample sizes to provide more power to explore relationships and secondary hypotheses, and the increased potential for improving the external validity of results by taking advantage of the heterogeneity among the studies in generalizing results to a wider context. Potential disadvantages are that combining different studies increases the overall variability, may produce spurious results, and could affect how the overall results are received by the scientific community. The heterogeneity among the studies can be such that it may actually reduce overall statistical power. In addition, conflicting results may make the overall result inconclusive, despite the analytic methodology, with wider confidence intervals due to the increased heterogeneity.

Thus, in deciding whether to combine data or not, the primary issues that must be considered include not only that they can address an important research question, but that the studies are

"sufficiently" comparable with respect to their conceptual framework and overall objectives as well as design and implementation features. The latter include participant eligibility criteria and characteristics; intervention settings, approach, components, timing and actual implementation; outcome(s) of interest (e.g. how and when measured); and study conduct procedures (e.g. staff training, quality assurance, data management). Some studies are implemented in a more 'pragmatic style' (flexibility in design characteristics), whereas others in a more 'explanatory style.' It is important to devise statistical tests that can inform the decisions for combining the data from RCTs.

Methods to summarize studies without producing a summary estimate

Systematic reviews are commonly conducted as a method for summarizing information from multiple studies that address the same or similar scientific questions. Combining the results from multiple similar randomized controlled trials to synthesize the empirical evidence related to a particular intervention is a well-established methodology in systematic reviews. Studies are required to meet strict pre-specified criteria to be included in such a review. The methods and results of each study are considered separately, but results are not necessarily combined quantitatively to produce a summary result in a systematic review. Often the commonalities and differences among studies are summarized in text or in table format. A systematic review conducted over studies within a consortium would highlight the common aspects of design, approach, data management and measurement that might be unique to that group of studies.

A comparative, as opposed to a summarizing, approach involves comparing the effect from one index study that was of particular interest to the effects found in other studies, one by one. This would provide a test of whether other studies corroborate (validate) the result of the index study. Although this does not provide an overall estimate of the intervention effect, it does provide insights into the underlying heterogeneity. It also prompts the investigator to search for reasons why some studies may be in agreement, while others are not.

A descriptive graphical approach that is useful for the comparison of studies is the 'forest plot'. In these plots, the estimate of the intervention effect and its corresponding 95% confidence interval are presented for each study as a line segment alongside each other. Studies may be arranged in

alphabetical order, chronologically, or by size of effect. Forest plots facilitate visual assessment of results from multiple studies, and can be used with or without the addition of a summary estimate of effect over the multiple studies.

Pooling methodologies to produce a combined estimate

Having made the decision to combine study estimates from multiple RCT's to produce a single estimate, several methods can be considered. Combining by collapsing all observations into a single data set and ignoring study differences is often referred to as 'lumping' the data. As illustrated by DeMets (1987), this approach may produce misleading results. For example, different interventions in different studies could produce strong results, but in opposite directions, resulting in the analysis of collapsed data showing a null effect. Combining by 'pooling' rather than lumping is preferable, with the term pooling meant to convey a method that statistically adjusts for the study differences. There are several alternatives for pooling, described below.

A specific methodology that may be employed if there are two interventions and the outcome variable is binomial is the Mantel-Haenszel (1959) method for combining data over several 2x2 contingency tables. For continuous variables, Mantel and Haenszel (1959) suggested ANOVA-based approaches for summarizing intervention effects across studies. Instead of ANOVA, one can choose the flexibility of regression models to incorporate study and intervention interaction effects by including appropriate indicator variables. These models can also include study-level and subject-level covariates if information is available on these levels (see Models #1a-c in the Appendix).

Meta-analysis is a well-known approach for obtaining a common intervention effect from several similar trials. The heterogeneity among the individual studies' estimates of effects, the within-study variance of the outcome measure(s), and a quality assessment of the studies, are determined. Combining widely disparate measures into a single summary measure masks conceivably important differences and is often discouraged. If the studies meet a pre-specified criterion of effect size homogeneity and other criteria for meaningful cross-study analyses, their individual results may be combined to produce an estimate of the intervention effect. A weighted pooled estimate is obtained, considering the inverse of each study's variance, under the assumption that the larger the

variance of a study, the lower the 'quality' of its evidence and therefore the less weight it should have upon the overall effect estimate. This variance may be calculated using either a fixed effects or a random effects approach. The random effects approach attenuates the variance estimates and thus the weights by considering within-study and among-study information (see further below).

However, these approaches, which assume the same or similar interventions for all active arm participants and all control arm participants, do not work for multi-component interventions that vary across sites, when in actuality the active and/or control arm subjects at one site may be receiving a different intervention than the active and/or control arm subjects at another site.

Another possible meta-analytic methodology is multiple intervention meta-analyses, e.g. network meta-analysis. It is used when there are not enough head-to-head comparisons of multiple interventions, and considers each randomized arm in calculating intervention effect estimates. It may be considered in situations where the same randomized arms are not included in all studies. In our situation, a given arm of a randomized study consists of an intervention with multiple components occurring simultaneously. Thus, the use of network meta-analysis, like the use of 'standard' meta-analysis, is not appropriate for multi-component interventions.

Table 1 presents a summary of the advantages and disadvantages of four common methods that can be used to address multi-component interventions: random effects meta analysis, meta-regression, multilevel meta-regression, a technique that includes individual participant-level and study-level data, and modeling of structural relationships.

**** Insert Table 1 about here****

Random-effects meta analysis

In meta-analysis, one is modeling the intervention effect, which is the same as modeling the expected value of the outcome in two-arm studies. In standard fixed-effects meta-analysis, the assumption is that there is a common intervention effect and that each observed study outcome effect differs from the true effect by an amount defined as the 'error term,' which is assumed to be normally distributed. If one is willing to assume that the studies are a random sample from a potential pool of all other similar studies, one can assume that each study's effect varies around its

own true study effect, thereby decomposing the total variance for the estimate of the intervention effect into a within-study variance and a between-study variance.

While the fixed-effects meta-analysis approach is widely used, it assumes that there is little heterogeneity in study effects across the various trials. The random effects approach differs from the fixed effects approach in that it considers heterogeneity information across the trials in calculating a trial's variance, while the fixed effects approach utilizes only within-study information for calculating a study's variance. Using a random effect approach in a meta-analysis does statistically adjust for some of the heterogeneity across studies. However, there may still be residual heterogeneity among the studies and meta-analytic techniques cannot account for multiple-component interventions.

Study-level meta regression

The technique of modeling the study level outcome by incorporating study-level covariate information is called meta-regression. To account for the additional or residual heterogeneity among the studies because of different intervention approaches or participant characteristics, one can model the outcome using study-level covariates and thus adjust for the effects of each study upon the outcome as well as upon the effect of the intervention (Bangdiwala et. al. 2012). A generalized linear mixed effects regression model on the primary study outcome is constructed, with an indicator variable for intervention arm, and study-level covariates that may be potential effect modifiers (moderators) of the intervention effect or potential confounders of the intervention effect. The introduction of study-level covariates in a meta-regression may explain some of the heterogeneity due to study differences (Morton et al 2004).

The issue of whether to include each study's effects as a random or fixed effect is not straightforward. DeMets (1987) cautioned against including random effects for the studies since this can imply that they are a random sample of a specified universe of studies. Moreover, one would require a large number of studies for estimating variances of the random effects for studies. It would thus seem appropriate to include study effects as fixed effects in the pooled estimation. On the other hand, including study effects as random variables accounts for the heterogeneity among the study effects due to the unobserved sources. Including study effects as random effects can be justified by the interest in adjusting for the studies' source of variability rather than in interpreting

those effects. Modeling random effects can lead to narrower confidence intervals around the estimates of intervention effectiveness (Bangdiwala et. al. 2012).

In order to understand the intervention effect when there are multiple components to the intervention, Bangdiwala et al (2012) proposed to include indicator variables for each component across the different studies. To avoid component effects being confounded with the study effect, it is important that a study use more than one component and that a particular component be used in more than one study. However, since it is likely that components are not exactly the same across studies, in order to consider components as 'similar', a common taxonomy could be utilized (O'Connor 2015) (see Tate et al in this issue). Note that the control or 'standard care' arm may include some 'active' components and they would also need to be accounted for in the analysis. Having fit a meta regression, one can then look at the coefficients of each of the component indicator variables to assess their relative contribution to the overall outcome. See Appendix Model #2 for an illustration using the COPTR consortium.

Multilevel meta-regression

Within consortia, investigators have the possibility of obtaining participant-level information in addition to study-level information. The latter might include various aspects of the interventions such as delivery characteristics, implementation strategies, and mechanisms of action (Schulz et al 2010). The meta-regression model can be expanded to include such information, and is then called multilevel meta-regression (see Model #3 in Appendix). Moderately large heterogeneity among the studies' target populations, intervention content and modalities and other aspects may be addressed using study-level along with participant-level covariates (Morton et al 2004).

In the Resources for Enhancing Alzheimer's Caregiver Health (REACH) consortium, this analytic approach was used to allow investigators to include in a single model both participant-level information and individual elements of multicomponent interventions at the study-level to examine the relationships between those elements and outcomes (Czaja et al 2003, Belle et al 2003). The REACH interventions were complex multi-faceted behavioral interventions, with various components. A natural question is which components are more effective, but since not all studies had the same components in their interventions, REACH investigators decomposed the complex

interventions into 12 components (e.g., caregiver affect, care-recipient behavior, knowledge about the social environment) and relationships between the components and outcome were examined. By so doing, main effects and interactions, both within levels (participant, study) and across those levels, were examined.

Modeling structural relationships

Multilevel meta regression models may account for the heterogeneity among studies and for the effects of the various components across studies, but fall short of considering the causal pathways, whether testing those mechanisms is an explicit objective or not. The paths are present in the overall framework for the study, which is why having a common 'framework' is crucial when pooling data. Whether those paths are measured and tested or not, they exist and affect the intervention impact. To the extent they can be modeled, they provide richer explanations for the variation in response.

Population-based interventions initially induce behavioral changes among the subjects that, in turn, affect the outcomes of interest (Bhargava, 2008). For example, making parents aware of the importance of healthy diets and greater physical activity for children, as is common in childhood obesity interventions, may lead to changes in parental behavior that in turn reduce childhood obesity. Similarly, highly motivated women in the Women's Health Trial: Feasibility Study in Minority Populations were seen to make healthful dietary changes especially in the intervention group (Bhargava and Hays, 2004). It is important to analyze the data from RCTs in a broad framework and investigate the pathways underlying the intervention effects. Moreover, exploratory analyses of pooled data from multi-center and/or similar trials can provide insights for the future design of effective interventions.

Multigroup structural equation modeling with means structures (MG-SEM) (Jöreskog 1971; Sörbom 1974) is an alternative to the regression approach that accommodates multiple components and pathways. The structural model is specified in each study population separately and common parameters are constrained to be equal across study groups. Lagrange multipliers are used to determine if constraints significantly worsen the model fit. When a constraint does not hold, parameters are estimated separately in each group. The study variables can be defined at the latent variable level by different combinations of observed variables and the differences in construct

reliability can be taken into account. The validity of latent constructs can be tested under certain identifying assumptions on variances of the variables. This methodology has been used in social sciences but is not common in the evaluation of clinical trials (Rabe-Hesketh, Skrondal & Pickles 2004; Duncan, Duncan, & Strycker, 2006). As discussed next, rigorous testing of the constancy of model parameters can also proceed in the regression framework by applying likelihood ratio tests and taking into account the unobserved between-subject differences via random effects.

Statistical tests for justifying pooling

The interpretations of treatment or intervention effects in randomized controlled trials can be complex (Fisher 1935, Cox, 1958). Pooling data from various studies may be useful for obtaining information that could not be gleaned from individual studies and can improve precision of estimates of intervention, or intervention component effects. However, it is important to apply likelihood ratio and other statistical tests for assessing the validity of the pooled estimates for avoiding potentially misleading inferences.

From the standpoint of rigorous justifications for pooling data from similarly designed RCT's, likelihood ratio statistics can be applied to test for the constancy of model parameters across sites (Bhargava and Guthrie, 2002). This is especially appealing in true multi-site trials and in situations where similar study designs are used for different population groups and relevant explanatory variables are available. For example, in the Women's Health Trial: Feasibility Study in Minority Populations, the effects of subjects' "unhealthy eating habits" and dietary intakes on body weight were likely to differ for Control and Intervention groups. By including separate intercept terms for Control and Intervention groups, the empirical models enabled testing of the null hypothesis that model parameters are the same for the two groups. In this case, the value of the likelihood ratio statistic was significant. (Bhargava and Guthrie, 2002). Of course, if the null hypothesis had not been rejected, it would have provided some justification for pooling the data for the two groups.

Further, the null hypothesis of constancy of model parameters may be rejected in certain applications via the use of likelihood ratio tests because the populations differ in important respects such as behavioral and socioeconomic aspects. In such circumstances, it would seem prudent not to

pool the data for increasing the sample sizes since that might entail increasing biases in the estimated parameters. However, as noted above, some *a priori* information can be incorporated in pooled analyses. For example, suppose that in a RCT the effect of an intervention is significant and the estimated model parameters indicate that an explanatory variable such as subjects' 'participation motivation' was associated with the changes. Then it may be useful to test if the coefficient of participation motivation does not differ statistically for other population groups for which smaller numbers of observations might be available. This null hypothesis can be tested using Lagrange Multiplier type tests (Rao, 1948) that require model estimation only under the null hypothesis. Moreover, Wald statistics can be applied to test the null hypothesis by estimating the model under the more general alternative hypothesis. In addition, likelihood ratio statistics are insightful since investigators can assess robustness of the estimated parameters under the null and alternative hypotheses (Sargan, 1980, Bhargava, 1987). Such statistical tests can be extended to situations where the errors may not be normally distributed though possess finite fourth order moments (Bhargava, 1987).

Conclusion

The question of whether to combine data across studies, such as may be seen in a consortium, does not have a simple answer. Difficult issues to consider are how to approach the problems and how to decide whether it will be useful to combine the data. Combining data from heterogeneous studies can lead to spurious results and conclusions. The argument of combining to achieve higher statistical power for the primary research hypotheses within a consortium of studies might be a weak one since each trial within a consortium is typically adequately powered to address those hypotheses. The potential for addressing the intervention effects within subgroups by pooling, for improving external validity, for asking research questions that are not possible to test in the individual studies such as examining intervention components, and for addressing secondary outcomes with increased power, may be quite attractive. If one decides to pool the data, heterogeneity among the studies' procedures and uniqueness of subject selection criteria and other important characteristics make it necessary to apply analytical and statistical tools that attempt to address these issues.

There are many potential methodologies for pooling data across studies (Weiner et al 2012). Whether weighting the different studies data differently, or via employing random effects, one must recognize that different pooling methodologies may yield different results. It is important to spell out the conceptual framework employed as well as the specific research questions for the pooled data *a priori* and apply appropriate statistical techniques.

As stated earlier, the objective of this manuscript is to describe certain issues to be considered when combining data across studies, especially studies that share many commonalities, as in consortia studies. In such situations, the number of studies is pre-determined and out of the control of investigators. In actual implementation of the methods described, the number of studies needed would be based on the desired precision for the actual estimation of effects and will depend on the variability of the outcome variable. The more studies the better the precision will be.

This manuscript is not proposing 'new' methods - but bringing them together in one place, to provide researchers with the advantages and the limitations of the currently available methodologies. The approaches presented here, whether modeling by random effects meta regression, or using multilevel structural equation models, involve adjusting for the increased heterogeneity in the data due to aggregating information across multiple studies. As for all data syntheses, the number of studies available for pooling is a consideration when using any of these techniques. In the modeling, it is possible to test for interaction and constancy of model parameters across studies in the pooled models via likelihood ratio and other tests. One can also set up sequential tests in certain cases where the hypotheses are nested (Wald 1947, Bhargava 1987).

In summary, pooling can be used for comprehensive exploratory analyses of data from RCT's and should not be viewed as replacing the standard analysis plan for each study. As noted above in the context of dietary interventions, pooling may help to identify new hypotheses about intervention components that may be more effective especially for subsets of participants with certain behavioral characteristics. Pooling, when supported by statistical tests, can allow exploratory investigation of interesting potential hypotheses and for the design of future interventions.

Acknowledgement

This manuscript is one of three presented in this journal and was supported the NIH National Heart, Lung, and Blood Institute, the *Eunice Kennedy Shriver* National Institute of Child Health and Development, the NIH Office of Behavioral and Social Sciences Research, the NIH Office of Disease Prevention, and the Centers for Disease Control and Prevention. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

All authors have completed the disclosure form on all relationships or interests that could influence or bias the work, and there are no conflicts of interests from any authors.

Adherence to Ethical Principles

All authors have adhered to ethical principles and maintain the integrity of the research and its presentation by following the rules of good scientific practice.

References

Bangdiwala SI, Villaveces A, Garrettson M, Ringwalt C (2012). Statistical methods for designing and assessing the effectiveness of community-based interventions with small numbers, *International Journal of Injury Control & Safety Promotion* 19(3):242-248.

Belle SH, Czaja SJ, Schulz R, Zhang S, Burgio LD, Gitlin LN, Jones R, Mendelsohn AB, Ory MG (2003). Using a new taxonomy to combine the uncombinable: Integrating results across diverse interventions, *Psychology and Aging* 18(3):396 -405.

Bhargava A (1987). Wald tests and systems of stochastic equations. *International Economic Review*, 28, 789-808.

Bhargava A. (2008). Randomized controlled experiments in health and social sciences: Some conceptual issues, *Economics and Human Biology*, 6:293-298.

Bhargava A, Guthrie J (2002). Unhealthy eating habits, physical exercise and macronutrient intakes are predictors of anthropometric indicators in the Women's Health Trial: Feasibility Study in Minority Populations, *British Journal of Nutrition* 88(6):719-28.

Bhargava A, Hays J (2004). Behavioral variables and education are predictors of dietary change in the Women's Health Trial: Feasibility Study in Minority Populations, *Preventive Medicine* 38(4):442-51.

Cox D (1958) *Planning of experiments*. New York: John Wiley & Sons.

Czaja SJ, Schulz R, Lee CC, Belle SH, Investigators REACH (2003) A methodology for describing and decomposing complex psychosocial and behavioral interventions, *Psychology and Aging* 18(3):385-395.

Czajkowski SM, Powell LH, Adler N, et al (2015) From ideas to efficacy: The ORBIT model for developing behavioral treatments for chronic diseases, *Health Psychology (to appear in print)*

DeMets DL (1987) Methods for combining randomized clinical trials: Strengths and limitations, *Statistics in Medicine* 6:341-348.

Duncan TE, Duncan SC, Strycker LA (2006) *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Application*, 2nd Ed., Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.

Fisher RA (1935) *The Design of Experiments*, Oliver and Boyd, Edinburgh.

Jöreskog KG (1971) Simultaneous factor analysis in several populations, *Psychometrika* 36:409-426.

Lytle LA, Svetkey LP, Patrick K, et al (2014) The EARLY trials: a consortium of studies targeting weight control in young adults, *Translational Behavioral Medicine* 4(3):304-313.

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* 22:719-748.

Morton SC, Adams JL, Suttorp MJ, Shekelle PG (2004) Meta-regression approaches: What, why, when, and how?, Technical Review 8, Agency for Healthcare Research and Quality Publication No. 04-0033. Rockville, MD.

O'Connor DP, Lee RE, Mehta P, Thompson D, Bhargava A, Carlson C, Kao D, Layne CS, Ledoux T, O'Connor T, Rifai H, Gulley L, Hallett AM, Kudia O, Joseph S, Modelska M, Ortega D, Parker N, Stevens A (2015) Childhood Obesity Research Demonstration project: cross-site evaluation methods, *Childhood Obesity* 11:92-103.

Pratt CA, Boyington J, Esposito L, et al (2013) Childhood Obesity Prevention and Treatment Research (COPTR): interventions addressing multiple influences in childhood and adolescent obesity, *Contemporary Clinical Trials* 36(2):406-413.

Rabe-Hesketh S, Skrondal A, Pickles A (2004) Generalized multilevel structural equation modeling, *Psychometrika* 69:167–190.

Rao CR (1948) Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, *Proceedings of the Cambridge Philosophical Society* 44:50-57.

Sargan JD (1980) Some tests of dynamic specification for a single equation, *Econometrica* 48:879-898.

Schulz R, Czaja SJ, McKay JR, Ory MG, Belle SH (2010) Intervention taxonomy (ITAX): describing essential features of interventions, *American Journal of Health Behavior* 34(6):811-821.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology* 27:229–239.

Spinks A, Turner C, Nixon J, McClure RJ (2009) The 'WHO Safe Communities' model for the prevention of injury in whole populations. *Cochrane Database of Systematic Reviews* Issue 3. Art. No.: CD004445. DOI: 10.1002/14651858.CD004445.pub3.

Wald A (1947) *Sequential analysis*, Dover Publications, New York.

Weiner BJ, Lewis MA, Clauser SB, Stitzenberg KB (2012) In search of synergy: Strategies for combining interventions at multiple levels, *Journal of the National Cancer Institute Monogr* 44:34-41.