NCEE 2008-4018

U. S. DEPARTMENT OF EDUCATION

Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences

Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations

May 2008

Peter Z. Schochet
Mathematica Policy Research, Inc.

Abstract

This report presents guidelines for addressing the multiple comparisons problem in impact evaluations in the education area. The problem occurs due to the large number of hypothesis tests that are typically conducted across outcomes and subgroups in these studies, which can lead to spurious statistically significant impact findings. The guidelines, which balance type I and type II errors, involve specifying confirmatory and exploratory analyses in the study protocols, structuring the data by delineating outcome domains, conducting t-tests on composite domain outcomes, and applying multiplicity correction procedures to composites across domains. Guidelines are discussed for subgroup analyses, designs with multiple treatment groups, power analyses, and reporting impact findings. The report also provides background for applying the guidelines, including a detailed discussion of the multiplicity problem, statistical solutions that are found in the literature, and weighting options for forming composite domain outcomes.

NCEE 2008-4018
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences under Contract ED-04-CO-0112/0006.

Disclaimer
The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Mathematica Policy Research, Inc. to develop guidelines for appropriately handling multiple testing in education research. These guidelines—which are presented in this report—were developed with substantial input from an advisory panel (Appendix A lists the panel members). The views expressed in this report, however, are those of the author and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

May 2008

Schochet, Peter Z. (2008). *Technical Methods Report*: *Guidelines for Multiple Testing in Impact Evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats
Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

The author for this report, Dr. Peter Schochet, is an employee of Mathematica Policy Research, Inc. with whom IES contracted to develop the guidelines that are presented in this report. Dr. Schochet and other MPR staff do not have financial interests that could be affected by the guidelines. In addition, no one on the thirteen-member Expert Advisory Panel, that convened three times to provide advice and guidance, has financial interests that could be affected by the report content.

Contents

List of Tables

Chapter One: Introduction

Studies that examine the impacts of education interventions on key student, teacher, and school outcomes typically collect data on large samples and on many outcomes. In analyzing these data, researchers typically conduct multiple hypothesis tests to address key impact evaluation questions. Tests are conducted to assess intervention effects for multiple outcomes, for multiple subgroups of schools or individuals, and sometimes across multiple treatment alternatives.

In such instances, *separate t*-tests for each contrast are often performed to test the null hypothesis of no impacts, where the Type I error rate (statistical significance level) is typically set at $\alpha = 5$ percent for each test. This means that, for each test, the chance of erroneously finding a statistically significant impact is 5 percent. However, when the hypothesis tests are considered *together*, the "combined" Type I error rate could be considerably larger than 5 percent. For example, if all null hypotheses are true, the chance of finding at least one spurious impact is 23 percent if 5 independent tests are conducted, 64 percent for 20 tests, and 92 percent for 50 tests (as discussed in more detail later in this report). Thus, without accounting for the multiple comparisons being conducted, users of the study findings may draw unwarranted conclusions.

At the same time, statistical procedures that correct for multiple testing typically result in hypothesis tests with reduced statistical power—the probability of rejecting the null hypothesis given that it is false. Stated differently, these adjustment methods reduce the likelihood of identifying real differences between the contrasted groups. This is because controlling for multiple testing involves lowering the Type I error rate for individual tests, with a resulting increase in the Type II error rate. Simulation results presented later in this report show that if statistical power for an uncorrected individual test is 80 percent, the commonly-used Bonferroni adjustment procedure reduces statistical power to 59 percent if 5 tests are conducted, 41 percent for 20 tests, and 31 percent for 50 tests. Thus, multiplicity adjustment procedures can lead to substantial losses in statistical power.

There is disagreement about the use of multiple testing procedures and the appropriate tradeoff between Type I error and statistical power (Type II error). Saville (1990) argues against multiplicity control to avoid statistical power losses, and that common sense and information from other sources should be used to protect against errors of interpretation. Cook and Farewell (1996) argue that multiplicity adjustments may not be necessary if there is a priori interest in estimating separate (marginal) treatment effects for a limited number of key contrasts that pertain to different aspects of the intervention. Some authors also contend that the use of multiplicity corrections may be somewhat ad hoc because the choice of the size and composition of the family tested could be "manipulated" to find statistical significance (or insignificance). Many other authors argue, however, that ignoring multiplicity can lead to serious misinterpretation of study findings and publishing bias (see, for example, Westfall et al. 1999). These authors argue also that the choice of the tested families should be made prior to the data analysis to avoid the manipulation of findings.

Multiple comparisons issues are often not addressed in impact evaluations of educational interventions or in other fields. For example, in a survey of physiology journals, Curran-Everett (2000) found that only 40 percent of articles reporting results from clinical trials addressed the multiple comparisons problem. Hsu (1996) reports also that multiple comparisons adjustment procedures are often used incorrectly.

Accordingly, the Institute of Education Sciences (IES) at the U.S. Department of Education (ED) contracted with Mathematica Policy Research, Inc. (MPR) to develop guidelines for appropriately handling multiple testing in education research. These guidelines—which are presented in this report—

were developed with substantial input from an advisory panel (Appendix A lists the panel members). The views expressed in this report, however, are those of the author.

The remainder of this report presents the guidelines for multiple testing, followed by several technical appendixes to help researchers apply the guidelines. Appendix B provides more details on the nature of the multiple testing problem and the statistical solutions that have been proposed in the literature. Appendix C discusses the creation of composite outcome measures, which is a central feature of the recommended procedures. Finally, Appendix D presents the Bayesian hypothesis testing approach, which is the main alternative to the classical hypothesis testing framework that is assumed for this report.

Chapter Two: Guidelines for Multiple Testing

This section first discusses basic principles for addressing multiplicity, followed by a presentation of testing strategy guidelines. The focus is on designs with a single treatment and control group where data on multiple outcomes are collected for each sample member. These are the most common designs that are used in IES-funded education research. Guidelines are also provided for subgroup analyses and for designs with multiple treatment groups. The guidelines are consistent with those proposed for medical trials (see, for example, Lang and Secic 2007, CPMP 2002, and Altman et al. 2001), but are designed for evaluations of education interventions.

This report provides a *structure* to address the multiplicity problem and discusses issues to consider when formulating a testing strategy. The report does not provide step-by-step instructions on how to apply the guidelines, which is not possible due to the myriad types of impact evaluations that are conducted in the education field. Specific details on the use of the guidelines will vary by study depending on the interventions being tested, target populations, key research questions, and study objectives.

Finally, the guidelines assume that a *classical* (frequentist) hypothesis testing approach is used to analyze the data, because this is the testing strategy that is typically used in impact evaluations in the education field. Appendix B discusses the basic features of this testing approach. (Appendix D discusses the alternative Bayesian approach.)

Basic Principles

1. The multiple comparisons problem should not be ignored.

The multiple comparisons problem can lead to erroneous study conclusions if the α level for individual tests is not adjusted downward. At the same time, strategies for dealing with multiplicity must strike a reasonable balance between testing rigor and statistical power—the chance of finding truly effective interventions.

2. Limiting the number of outcomes and subgroups forces a sharp focus and is one of the best ways to address the multiple comparisons problem.

Multiple testing is less of a problem if studies limit the number of contrasts for analysis. Sharply focusing research questions on one or a few outcomes and on a small number of target groups diminishes the chance of finding impacts where none exist.

At the same time, in some studies, theory and prior research may not support a sharp focus on outcomes or subgroups and, in others, the tested interventions may be expected to have a range of effects. Furthermore, in a context where IES and other funders are executing costly studies to identify promising approaches to difficult problems, narrowing the range of outcomes and subgroups limits researchers' ability to use post hoc exploratory analyses to find unexpected, yet policy-relevant information.

Thus, the multiple comparisons testing strategy should be flexible to allow for (1) *confirmatory* analyses to assess how strongly the study's pre-specified central hypotheses are supported by the data, and (2) *exploratory* analyses to identify hypotheses that could be subject to future rigorous testing.

3. The multiple comparisons problem should be addressed by first structuring the data. Furthermore, protocols for addressing the multiple comparisons problem should be made *before* data analysis is undertaken.

The multiple comparison testing strategy should be based on a process that first groups and prioritizes outcomes. The structuring of the data should be specified during the design stage of the study and published before study data are collected and analyzed. Multiple comparisons corrections should *not* be applied blindly to all outcomes, subgroups, and treatment alternatives considered together. This approach would produce unnecessarily large reductions in the statistical power of the tests. Rather, the testing strategy should strike a reasonable balance between testing rigor and statistical power.

Specific plans for structuring the data and addressing the multiple comparisons issue will depend on the study objectives. However, the testing strategy described next pertains broadly to impact evaluations that are typically conducted in the education field.

Guidelines for Developing a Strategy for Multiple Testing

1. Delineate separate outcome domains in the study protocols.

Outcome domains should be delineated using theory or a conceptual framework that relates the program or intervention to the outcomes. The domains should reflect key clusters of constructs represented by the central research questions of the study.

The outcome domains, for example, could be defined by grouping outcomes that are deemed to have a common latent structure (such as test scores in particular subject areas, behavioral outcomes, or measures of classroom practices) or grouping outcomes with high correlations. Domains could also be defined for the same outcomes measured over time (for example, test scores collected at various follow-up points). Domains could pertain to specific population subgroups (for example, if the intervention is targeted primarily to students with particular characteristics, such as English language learners).

2. Define confirmatory and exploratory analysis components prior to data analysis.

The *confirmatory* analysis should provide estimates whose statistical properties can be stated precisely. The goal of this analysis is to present rigorous tests of the study's central hypotheses that are specified in the study protocols. The confirmatory analysis *must* address multiple comparison issues and must have sufficient statistical power to address the main research questions. This analysis could consist of two parts: (1) testing for impacts for each outcome domain separately, and (2) jointly testing for impacts across outcome domains. These analyses do not necessarily need to include all domains.

The purpose of the *exploratory* analysis is to examine relationships within the data to identify outcomes or subgroups for which impacts may exist. The goal of the exploratory analysis is to identify hypotheses that could be subject to more rigorous future examination, but cannot be examined in the present study because they were not identified ahead of time or statistical power was deemed to be insufficient. Results from post hoc analyses are not automatically invalid, but, irrespective of plausibility or statistical significance, they should be regarded as preliminary and unreliable unless they can be rigorously tested and replicated in future studies.

3. For *domain-specific* confirmatory analyses, conduct hypothesis testing for domain outcomes as a group.

Outcomes will likely be grouped into a domain if they are expected to measure a common latent construct (even if the precise psychometric properties of the domain "items" are not always known in advance). Thus, conducting tests for domain outcomes as a group will measure intervention effects on this common construct. Combining outcomes that each tap the same latent construct could also yield test statistics with greater statistical power than if individual outcomes were examined one at a time.

A *composite t-test* approach is recommended for testing global hypotheses about a domain. Under this approach, significance testing is performed on a single combination of domain outcomes. This procedure accounts for multiple comparisons by reducing the domain outcome to a single composite measure. This approach addresses the question "Relative to the status quo, did the intervention have a statistically significant effect on a typical domain outcome or common domain latent factor?" Appendix C discusses possible options for defining weights to construct composite outcome measures.

A statistically significant finding for a composite measure provides confirmatory evidence that the intervention had an effect on the common domain latent construct. When statistical significance of the composite has been established, a within-domain exploratory analysis could be conducted using *unadjusted p*-values to identify specific domain outcomes that contributed to the significant overall effect. The significance of a particular outcome does *not* provide confirmatory evidence about the domain as a whole, but provides information that could be used to help interpret the global findings.

If the impact on the composite measure is not statistically significant, it is generally not appropriate to examine the statistical significance of the individual domain outcomes. However, if such analyses are performed, they must be qualified as exploratory.

4. Use a similar testing strategy if the confirmatory analysis involves assessing intervention effects *across* domains.

Providing confirmatory evidence about intervention effects for each domain separately may satisfy research objectives in some studies. However, if an intervention is to be judged based on its success in improving outcomes in one or more domains, the study may wish to obtain confirmatory summative evidence about intervention effects across domains. For instance, if test score and school attendance outcomes are delineated into separate domains, it may be of interest to rigorously test whether the intervention improved outcomes in either domain (or in both domains). In these cases, the study may wish to conduct hypothesis tests when the domains are considered together.

The appropriate use of multiplicity adjustments for such analyses will depend on the main research questions for assessing overall intervention effects. For example, the research question of interest may be "Did the intervention have an effect on *each* domain?" In this case, multiple comparisons corrections are not needed; separate *t*-tests should be conducted at the α significance level for each domain composite outcome, and the null hypothesis of no treatment effect in at least one domain would be rejected if each composite impact is statistically significant. This approach applies a very strict standard for confirming intervention effects.

The research question could instead be "Did the intervention have an effect on *any* domain?" In this case, multiplicity corrections are warranted. Different domains will likely tap different underlying latent factors and dimensions of intervention effects. Thus, rather than conducting a *t*-test on an aggregate composite measure across domains (which could be difficult to interpret), hypothesis testing could be conducted for each domain composite *individually* using the recommended statistical adjustment procedures discussed

in Appendix B. The null hypothesis of no treatment effect in each domain would then be rejected if any domain impact is statistically significant after applying the adjustment procedures.

5. Multiplicity adjustments are not required for exploratory analyses.

For exploratory analyses, one approach is to conduct unadjusted tests at the α level of significance. However, to minimize the chance of obtaining spurious significant findings, researchers may wish to apply multiplicity adjustments for some exploratory analyses if the data can be structured appropriately and statistical power levels are deemed to be tolerable.

Study reports should explicitly state that exploratory analyses do not provide rigorous evidence of the intervention's overall effectiveness. Results from post hoc analyses should be reported as providing preliminary information on relationships in the data that could be subject to more rigorous future examination. These qualifications apply *even if* multiple comparisons correction procedures are used for exploratory analyses.

6. Specify which *subgroups* will be part of the confirmatory analysis and which ones will be part of the exploratory analysis.

If the study seeks to make rigorous claims about intervention effects for specific population subgroups, this should be specified in the study protocols and embedded in the confirmatory analysis strategy. To limit the multiple testing problem, only a limited number of educationally meaningful subgroups should be included in the analysis. The direction of the expected subgroup effects should be specified and the exact subgroup definitions should be defined explicitly at the outset to avoid post hoc data-dependent definitions. Furthermore, to ensure treatment-control group balance for each subgroup, efforts should be made to conduct random assignment within strata defined by the subgroups. The case for including subgroups in the confirmatory analysis is stronger if there is a priori reason to believe that treatment effects differ across the subgroups.

The testing strategy for the subgroup analysis needs to address the simultaneous testing of multiple subgroups and outcomes and should link to the key study research questions. For example, suppose that hypotheses are postulated about intervention effects on a composite outcome across gender and age subgroups. Suppose also that the two key research questions are (1) "Is the intervention more effective for boys than girls?" and (2) "Is the intervention more effective for older than younger students?" In this case, it is appropriate to examine whether intervention effects *differ* by gender (age) by conducting *F*-tests on treatment-by-subgroup interaction terms that are included in the regression models. If the gender and age subgroups are to be considered together, a multiple comparisons adjustment procedure could be applied to the *p*-values from the various subgroup tests (see Appendix B).

Hypothesis tests of differential subgroup effects are appropriate if subgroup results are to be used to target future program services to specific students. In this case, the standard of confirmatory evidence about the subgroup findings should be set high. It is generally accepted in the medical literature that tests of interactions are more appropriate for subgroup analyses than separate, subgroup-specific analyses of treatment effects, because declarations of statistical significance are often associated with decision making (Brookes et al. 2001, Rothwell 2005, Gelman and Stern 2006).

There may be instances in education research, however, where the key research question is "Did the intervention have an effect on a composite outcome for a specific subgroup *in isolation*?" In this case, multiplicity adjustments are not warranted, because program effects are to be examined for a single subgroup that is specified in advance. Multiplicity adjustments, however, are necessary for hypothesis tests that address the following research questions: "Did the intervention have an effect for *either* younger

or older students?" or "Did the intervention have an effect for *any* gender or age subgroup?" Results from these tests, however, must *not* be interpreted as providing information about differential effects across subgroups.

Impact findings for subgroups that are not part of the confirmatory analysis should be treated as exploratory. Furthermore, post hoc subgroup analyses must be qualified as such in the study reports.

7. Apply multiplicity adjustments in experimental designs with *multiple* treatment groups.

A rigorous standard of evidence should be applied in designs with multiple treatment groups before concluding that some (for example, specific reading curricula) are preferred over others. The confirmatory testing strategy for these designs must be specified prior to data analysis. The strategy could include global tests of differences across treatments, or tests of differences between specific treatment pairs that could be used to rank treatments. The strategy should also address simultaneous testing of multiple treatments, outcomes, and subgroups (if pertinent). As discussed in Appendix B, multiplicity adjustment procedures have been developed for situations where multiple treatments are compared to each other or to a common control group.

8. Design the evaluation to have *sufficient statistical power* for examining intervention effects for all prespecified confirmatory analyses.

Statistical power calculations for the confirmatory analysis must account for multiplicity. The determination of appropriate evaluation sample sizes will depend on the nature of the confirmatory analysis. For example, for domain-specific confirmatory analyses, the study should have sufficient power to detect impacts for composite domain outcomes. Similarly, if subgroup analyses are part of the confirmatory testing strategy, the power analysis should account for the simultaneous testing of multiple subgroups and multiple outcomes, and similarly for designs with multiple treatment groups. Brookes et al. (2001) show that if a study has 80 percent power to detect the overall treatment effect, the sample needs to be at least four times larger to detect a subgroup-by-treatment interaction effect of the same magnitude.

9. *Qualify* confirmatory and exploratory analysis findings in the study reports.

There is no one way to present *p*-values from the multiple comparisons tests that will fit the needs of all evaluation reports. In some instances, it may be preferable to present adjusted *p*-values in appendices and the unadjusted *p*-values in the main text, whereas in other instances, it may be preferable to present adjusted *p*-values in the main text or in footnotes. The reporting of adjusted or unadjusted confidence intervals could also be desirable.

Some users of study reports may have a narrow interest in the effectiveness of the intervention for a specific outcome, subgroup, or treatment alternative. Where interest focuses on a specific contrast in isolation, the usual *t*-test conducted at significance level α is the appropriate test (and unadjusted *p*-values should be examined to assess statistical significance). This does not necessarily mean, however, that unadjusted *p*-values should be reported for all analyses to accommodate readers with myriad interests. Rather, study results should be presented in a way that best addresses the key research questions specified in the study protocols.

It is essential that results from the confirmatory and exploratory analyses be interpreted and qualified appropriately and that the presentation of results be consistent with study protocols. Confirmatory analysis findings should be highlighted and emphasized in the executive summary of study reports.

Appendix A

Panel Members Attending the Multiple Comparisons Meetings

Panel Members	Affiliation	Attended 2007 Meeting		
		February	July	December
Chairs				
Phoebe Cottingham	IES	✓	✓	✓
Rob Hollister	Swarthmore College	✓	✓	✓
Rebecca Maynard	University of Pennsylvania	✓	✓	
Participants				
Steve Bell	Abt Associates	✓	✓	✓
Howard Bloom	MDRC	✓	✓	✓
John Burghardt	Mathematica Policy Research, Inc.	✓	✓	✓
Mark Dynarski	Mathematica Policy Research, Inc.	✓	✓	✓
Andrew Gelman	Columbia University	✓		
David Judkins	Westat	✓	✓	✓
Jeff Kling	Brookings Institution	✓		✓
David Myers	American Institutes for Research	✓	✓	
Larry Orr	Abt Associates	✓	✓	✓
Peter Schochet	Mathematica Policy Research, Inc.	✓	✓	✓

Appendix B

Introduction to Multiple Testing

This appendix introduces the hypothesis testing framework for this report, the multiple testing problem, statistical methods to adjust for multiplicity, and some concerns that have been raised about these solutions. The goal is to provide an intuitive, nontechnical discussion of key issues related to this complex topic to help education researchers apply the guidelines presented in the report. A comprehensive review of the extensive literature in this area is beyond the scope of this introductory discussion. The focus is on continuous outcomes, but appropriate procedures are highlighted for other types of outcomes (such as binary outcomes). The appendix concludes with recommended methods.[1]

The Hypothesis Testing Framework

In this report, it has been assumed that a *classical* (frequentist) hypothesis testing approach is used to analyze the data; this is the testing strategy that is typically used in IES evaluations. This section highlights key features of this approach. Appendix D summarizes key features of the alternative Bayesian testing approach.

To describe the classical approach, it is assumed that treatment and control groups are randomly selected from a known population and that data on multiple outcomes are collected on each sample member. For contrast j, let μ_{Tj} and μ_{Cj} be population means for the treatment and control groups (or two treatment groups), respectively, and let $\delta_j = \mu_{Tj} - \mu_{Cj}$ be the population average treatment effect (impact). In the classical framework, population means and, hence, population impacts are assumed to be fixed.

Statistical analysis under this approach usually centers on a significance test—such as a two-tailed t-test—of a null hypothesis H_{0j}: $\delta_j = 0$ versus the alternative hypothesis H_{1j}: $\delta_j \neq 0$.[2] The Type I error rate—the probability of rejecting H_{0j} given that it is true—is typically set at $\alpha = 5$ percent for each test. Evaluation sample sizes are typically determined so that statistical power—the probability of rejecting H_{0j} given that it is false—is 80 percent if the true impact is equal to a value that is deemed to be educationally meaningful or realistically attainable by the intervention.

Under this framework, a null hypothesis is typically rejected (that is, an impact is declared statistically significant) if the p-value for the statistical test is less than 5 percent, or equivalently, if the 95 percent confidence interval for the contrast does not contain zero. This is a frequentist approach, because in hypothetical repeated sampling of population subjects to the treatment and control groups, 95 percent of the constructed confidence intervals would contain the true, fixed population impact. Probabilistic statements can be made about the random confidence interval but *not* about the fixed impact.

---

[1] Kirk (1994) provides an excellent introduction to the basics of statistical inference and multiple testing. Westfall et al. (1999), Shaffer (1995), and Hsu (1996) are excellent sources for more detailed discussions of the multiple comparisons problem. Savitz and Olshan (1995) discuss the multiple comparisons issue in the context of interpreting epidemiologic data.

[2] Under a one-tailed test, the alternative hypothesis is H_{1i}: $\delta_j > 0$ if larger values of the outcome are desirable or H_{1j}: $\delta_j < 0$ if smaller values of the outcome are desirable.

What Is the Multiple Testing Problem?

Researchers typically perform many simultaneous hypothesis tests when analyzing experimental data. Multiple tests are conducted to assess intervention effects (treatment-control differences) across multiple outcomes (endpoints). In some evaluations, multiple tests are also conducted to assess differences in intervention effects across multiple treatment groups (such as those defined by various reading or math curricula) or population subgroups (such as student subgroups defined by age, gender, race/ethnicity, or baseline risk factors).

In such instances, *separate t*-tests for each contrast are often performed to test the null hypothesis of no impacts, where the Type I error rate is typically set at $\alpha = 5$ percent for each test. Thus, for each test, the chance of erroneously finding a statistically significant impact is 5 percent. However, when the "family" of hypothesis tests are considered *together*, the "combined" Type I error rate could be considerably larger than 5 percent. This is the heart of the multiple testing problem.

For example, suppose that the null hypothesis is true for each test and that the tests are independent. Then, the chance of finding at least one spurious impact is $1 - (1 - \alpha)^N$, where N is the number of tests. Thus, the probability of making at least one Type I error is 23 percent if 5 tests are conducted, 64 percent for 20 tests, and 92 percent for 50 tests (Table B.1).

Table B.1: Chances of Findings Spurious Impacts for Independent Tests

Number of Independent Tests With True Null Hypotheses	Probability That at Least One *t*-Test Is Statistically Significant
5	0.23
10	0.40
20	0.64
50	0.92

The definition of the combined Type I error rate has implications for the strategies used to adjust for multiple testing and for interpreting the impact findings. The next section discusses the most common definitions found in the literature and provides numerical examples.

Definitions of the Combined Type I Error Rate

The two most common definitions of the combined Type I error rate found in the literature are the (1) *familywise error rate* (FWER) and (2) *false discovery rate* (FDR):

- The FWER, defined by Tukey (1953), has traditionally been the focus of research in this area. The FWER is the probability that at least one null hypothesis will be rejected when all null hypotheses are true. For example, when testing treatment-control differences across multiple outcomes, the FWER is the likelihood that at least one impact will be found to be significant when, in fact, the intervention had no effect on any outcome. As discussed, the FWER is $1 - (1 - \alpha)^N$ for independent tests, where N is the number of tests (dependent tests are discussed below).

- The FDR, defined by Benjamini and Hochberg (1995), is a more recent approach for assessing how errors in multiple testing could be considered. The FDR is the expected proportion of all rejected null hypotheses that are rejected erroneously. Stated differently, the FDR is the expected fraction of significant test statistics that are false discoveries.

Table B.2 helps clarify these two error rates. Suppose that multiple tests are conducted to assess intervention effects on N study outcomes and that M null hypotheses are true (M is unobservable). Suppose further that based on t-tests, Q null hypotheses are rejected and that A, B, C, and D signify cell counts when t-test results are compared to the truth. The counts Q and A to D are random variables.

Table B.2: The Number of Errors When Testing Multiple Hypotheses

	Results from Hypothesis Tests (Observed)		
Truth (Unobserved)	**H_{0j} Is Not Rejected**	**H_{0j} Is Rejected**	**Total**
H_{0j} Is True (No Impact)	A	B	M
H_{0j} Is False (Beneficial or Harmful Impacts)	C	D	$(N - M)$
Total	$(N - Q)$	Q	N

In Table B.2, the FWER is the probability that the random variable B is at least 1 among the M null hypotheses that are true. The FDR equals the expected value of B/Q, where B/Q is defined to equal 0 if $Q = 0$.[3] If all null hypotheses are true, then $B = Q$ and the FDR and FWER are equivalent, otherwise the FDR is smaller than or equal to the FWER.

The two error rates have a different philosophical basis. The FWER measures the likelihood of a *single* erroneous rejection of the null hypothesis across the family of tests. Researchers who focus on the FWER are concerned with mistakenly reporting *any* statistically significant findings. The concerns are that unwarranted scientific conclusions about evaluation findings could be made as a result of even one mistake and that researchers may select erroneous significant findings for emphasis when reporting and publishing results.

The rationale behind the FDR is that a few erroneous rejections may not be as problematic for drawing conclusions about the family tested when many null hypotheses are rejected as they would be if only a few null hypotheses are rejected. The rejection of many null hypotheses is a signal that there are real differences across the contrasted groups. Thus, researchers might be willing to tolerate *more* false positives (that is, larger values for B) if realized values for Q were large than if they were small. Under this approach, conclusions regarding intervention effects are to be based on the preponderance of evidence; the set of discoveries is to be used to reach an overall decision about the treatment. For those who adopt this approach, controlling the FWER is too conservative because, if many significant effects are found, a few additional errors will not change the overall validity of study findings.

[3] Mathematically, the FDR equals $E(B/Q \mid Q > 0)P(Q > 0)$.

Finally, variants of the FWER and FDR have been proposed in the literature. For example, Gordon et al. (2007) discuss a variant of the FWER—the per family error rate (PFER)—which is the expected *number* of Type I errors that are made (that is, the expected value of B in Table B.2). For instance, the PFER equals 5 if 5 of all tests with true null hypotheses are expected to be statistically significant. Gordon et al. (2007) argue that the use of the PFER (which focuses on expectations) rather than the FWER (which focuses on probabilities) could yield tests with more statistical power while maintaining stringent standards for Type I error rates across the family of tests. Storey (2002) introduced a variant of the FDR—the positive false discovery rate (pFDR)—which is the expected value of B/Q given that Q is positive. He argues that this measure may be more appropriate in some instances.

Quantifying the FWER and FDR

To demonstrate the relationship between the FDR and the FWER, simulated data were generated on N mean outcomes for samples of 1,000 treatment and 1,000 control group members (Table B.3). Data for mean outcome j were obtained as random draws from a normal distribution with standard deviation 1 and with mean μ_{Cj} for controls and μ_{Tj} for treatments. The impacts ($\mu_{Tj} - \mu_{Cj}$) were set to 0 for M outcomes with true null hypotheses and to 0.125 for ($N - M$) outcomes with false null hypotheses; the 0.125 value was set so that the statistical power of the tests was 80 percent. Each individual hypothesis was tested using a two-tailed t-test at the 5 percent significance level and the test statistics were generated independently. Each simulation involved 10,000 repetitions. Simulations were conducted for N = 5, 10, 20, and 50 and for M/N = 100, 80, 50, and 20 percent.

Table B.3: FWER and FDR Values for Independent Tests

Number of Tests	Percentage of Tests with True Null Hypotheses (No Impacts)	FWER	FDR
5	100	0.23	0.23
10	100	0.40	0.40
20	100	0.64	0.64
50	100	0.92	0.92
5	80	0.19	0.12
10	80	0.34	0.15
20	80	0.56	0.17
50	80	0.87	0.19
5	50	0.12	0.05
10	50	0.23	0.05
20	50	0.40	0.05
50	50	0.72	0.05
5	20	0.05	0.02
10	20	0.10	0.02
20	20	0.19	0.02
50	20	0.40	0.02

Note: FWER and FDR values were calculated using simulated data as described in the text.

The main results in Table B.3 are as follows:

- ***The FWER increases substantially with the number of tests.*** If all null hypotheses are true, the FWER is 23 percent for 5 independent tests, 64 percent for 20 independent tests, and 92 percent for 50 independent tests.
- ***The FWER and FDR are equivalent if all null hypotheses are true; otherwise, the FDR is less than the FWER.*** Thus, procedures that control the FWER also control the FDR, but the reverse does not necessarily hold. Differences between the FDR and FWER become larger as (1) the number of tests increases and (2) the number of true null hypotheses decreases (that is, when many differences across the contrasted groups truly exist).

These results suggest that the FDR is a less conservative measure than the FWER, especially if a considerable fraction of all null hypotheses are false. Thus, as demonstrated below, methods that control the FDR could yield tests with greater statistical power than those that control the FWER. The choice of which error criterion to control is important and must be made prior to the data analysis.

What Are Statistical Solutions to the Multiple Testing Problem?

A large body of literature describes statistical methods to adjust Type I errors for multiple testing (see, for example, the books by Westfall et al. 1999, Hsu 1996, and Westfall and Young 1993). The literature suggests that there is not one method that is preferred in all instances. Rather, the appropriate measure will depend on the study design, the primary research questions that are to be addressed, and the strength of inferences that are required.

This section briefly summarizes the literature in this area. Methods that control the FWER are discussed first, and methods that control the FDR are discussed second. Statistical packages (such as SAS) can be used to apply many of these methods.

Methods for FWER Control

Until recently, most of the literature on multiple testing focused on methods to control the FWER at a given α level (that is, methods to ensure that the FWER $\leq \alpha$). The most well-known method is the Bonferroni procedure, which sets the significance level for individual tests at α/N, where N is the number of tests.

The Bonferroni procedure controls the FWER when all null hypotheses are true or when some are true and some are false (that is, it provides "strong" control of the FWER). This feature differs from another well-known procedure—Fisher's protected least significant difference (LSD)—where an overall F-test across the tests is first conducted at the α level and further comparisons about individual contrasts are conducted at the α level only if the F-test is significant (Fisher 1935). Fisher's LSD controls the FWER only when all null hypotheses are true and, thus, provides "weak" control of the FWER. This means that Fisher's LSD may *not* control the FWER for second-stage individual hypotheses.[4] The same issue applies to other multiple-stage tests, such as the Newman-Keuls (Newman 1939, Keuls 1952) and Duncan (1955) methods.

[4] For example, suppose that there are 10 tests and that the null hypothesis is true for 9 tests. Suppose also that the true contrast for the 10th test is so large that the null hypothesis for the composite F-test would always be rejected. In this case, the Type I error rate would not be controlled for the second-stage t-tests, because there would be a 37 percent chance that these second-stage tests would reject at least one of the 9 tests with true null hypotheses.

The Bonferroni method applies to both continuous and discrete data, controls the FWER when the tests are correlated, and provides adjusted confidence bounds (by using α/N rather than α in the calculations). Furthermore, it is flexible because it controls the FWER for tests of joint hypotheses about *any* subset of N separate hypotheses (including individual contrasts). The procedure will reject a joint hypothesis H_0 if *any* p-value for the individual hypotheses included in H_0 is less than α/N. The Bonferroni method, however, yields conservative bounds on Type I error and, hence, has low power.

Many modified and sometimes more powerful versions of the Bonferroni method have been developed that provide strong control of the FWER. We provide several examples:

- Šidák (1967) developed a slightly less conservative bound where the significance level for individual tests is set at $1 - (1 - \alpha)^{1/N}$ rather than α/N. This method has properties similar to those of the Bonferroni method and is slightly more powerful, although it does not control the FWER in all situations in which test statistics are dependent.

- Scheffé (1959) developed an alternative procedure where two means are declared significantly different if $|t| \geq \sqrt{(N-1)F(\alpha; N-1, \nu)}$, where t is the t-statistic and $F(.)$ is the α-level critical value of the F distribution with $(N - 1)$ numerator and ν denominator degrees of freedom. This procedure has the nice property that if the F-test for the global hypothesis is insignificant, then the Scheffé method will never find any mean difference to be significant. The procedure applies also to all linear combinations of contrasts. It tends to be more powerful than the Bonferroni method if the number of tested contrasts is large (more than 20), but tends to be less powerful than the Bonferroni method for fewer tests.

- Holm (1979) developed a sequential "step-down" method: (1) order the p-values from the individual tests from smallest to largest, $p_{(1)} \leq p_{(2)} \ldots \leq p_{(N)}$, and order the corresponding null hypotheses $H_{0(1)}, H_{0(2),} \ldots, H_{0(N)}$; (2) define k as the minimum j such that $p_{(j)} > \alpha / (N - j + 1)$; and (3) reject all $H_{0(j)}$ for $j = 1, \ldots, (k - 1)$. This procedure is more powerful than the Bonferroni method because the bound for this method sequentially increases whereas the Bonferroni bound remains fixed. The Holm method controls the FWER in the strong sense, but cannot be used to obtain confidence intervals.

- Hochberg (1988) developed a "step-up" procedure that involves sequential testing where p-values are ordered from largest to smallest (rather than vice versa as for the Holm test). The method first defines k as the maximum j such that $p_{(j)} \leq \alpha/(N - j + 1)$, and then rejects all $H_{0(j)}$ for $j = 1, .., k$. This procedure is more powerful than the Holm method, but the control of the FWER is not guaranteed for all situations in which the test statistics are dependent (although simulation studies have shown that it is conservative under many dependency structures).

- Rom (1990) derived a step-up procedure similar to Hochberg's procedure that uses different cutoffs and has slightly more power because it *exactly* controls the FWER at α for independent test statistics.

Bootstrap and permutation resampling methods are alternative, computer-intensive methods that provide strong control of the FWER (see, for example, Westfall and Young 1993 and Westfall et al. 1990). These methods incorporate distributional and correlational structures across tests, so they tend to be less conservative than the other general-purpose methods and, hence, may have more power. Furthermore, they are applicable in many testing situations. These methods can be applied as follows:

- Generate a large number of pseudo data sets by selecting observations with replacement (for the bootstrap methods) or without replacement (for the permutation methods). The sampling

should be performed *without* regard to treatment status. Instead, sampling is performed for the combined research groups, and sampled observations are randomly ordered and proportionately split into pseudo-research groups.

- For each iteration, calculate pseudo-*p*-values for each *t*-test and store the minimum pseudo-*p*-value across tests.
- The adjusted *p*-value for an individual contrast is the proportion of iterations where the minimum pseudo-*p*-value is less than or equal to the *actual p*-value for that contrast.
- Significance testing is based on the adjusted *p*-values.

The intuition behind this procedure is that the distribution of the *maximum t*-statistic (*minimum p*-value) provides simultaneous confidence intervals that apply to all tests under the null hypothesis. The resampling methods use the data to estimate this distribution, which yields the multiplicity-adjusted *p*-values. In essence, a hypothesis test is rejected if the *actual t*-statistic value for that test is in the tail of the maximum *t*-statistic distribution.

Alternative methods to control the FWER have been developed when the design contains several treatment and control groups. The Tukey-Kramer (Tukey 1953, Kramer 1956) method is applicable if *all* pairwise comparisons of treatment and control means are of primary interest. If comparisons with a single control group are of primary interest, the Dunnett (1955) method is appropriate. These methods account for the dependence across test statistics due to the repetition of samples across contrasts.

The use of planned orthogonal contrasts is another method that adjusts for dependency when T treatment and control groups are compared to each other (see, for example, Bechhofer and Dunnett 1982). To describe this procedure, let $\overline{Y}_i$ be a mean outcome (composite) for research group i and let $C_j = \sum_{i=1}^{T} c_{ji}\overline{Y}_i$, where c_{ji} are constants such that $\sum_{i=1}^{T} c_{ji} = 0$ $(j = 1,...,(T-1))$. The C_js represent a family of $(T-1)$ contrasts (linear combinations) of the $\overline{Y}_i$ s. Mutually orthogonal contrasts arise if sample sizes are the same in each treatment condition and $\sum_{i=1}^{T} c_{ji}c_{ki} = 0$ for all $j \neq k$. A property of orthogonal contrasts is that the total sum of squares across the T research groups can be partitioned into $(T-1)$ sums of squares for each orthogonal contrast.

Significance testing can be performed for each orthogonal contrast using the multiple comparisons adjustment procedures discussed above. An advantage of this method is that testing problems associated with dependent test statistics disappear. Furthermore, if T is large, the use of orthogonal contrasts requires fewer test statistics than the Tukey-Kramer procedure, thereby reducing the multiple comparisons problem.

The use of planned orthogonal contrasts may be desirable if they correspond to key research questions and can be easily interpreted. However, this approach may not be appropriate if the key contrasts of interest are not mutually orthogonal.

Finally, special tests for controlling the FWER have been developed for binary or discrete data (see, for example, Westfall et al. 1999). Resampling methods or a modified Bonferroni method (Westfall and Wolfinger 1997) can be used in these instances. An alternative is the Freeman-Tukey Double Arcsine Test (Freeman and Tukey 1950).

Methods for FDR Control

Benjamini and Hochberg (1995) showed that when conducting N tests, the following four-step procedure will control the FDR at the α level:

1. Conduct N separate t-tests, each at the common significance level α.
2. Order the p-values of the N tests from smallest to largest, where $p_{(1)} \le p_{(2)} \le \ldots \le p_{(N)}$ are the ordered p-values.
3. Define k as the maximum j for which $p_{(j)} \le \frac{j}{N}\alpha$.
4. Reject all null hypotheses $H_{o(j)}$ $j = 1, 2, \ldots, k$. If no such k exists, then no hypotheses are rejected.

This "step-up" sequential procedure, which has become increasingly popular in the literature, is easy to use because it is based solely on p-values from the individual tests. Benjamini and Hochberg (1995) first proved that this procedure—which is hereafter referred to as the BH procedure—controls the FDR for continuous test statistics and Benjamini and Yekutieli (2001) proved that this procedure also controls the FDR for discrete test statistics.

The original result in Benjamini and Hochberg (1995) was proved assuming independent tests corresponding to the *true* null hypotheses (although independence was not required for test statistics corresponding to the *false* null hypotheses). Benjamini and Yekutieli (2001) proved, however, that the BH procedure also controls the FDR for true null hypotheses with "positive regression dependence." This technical condition is satisfied for some test statistics of interest, such as one-sided multivariate normal tests with nonnegative correlations between tests, but is not satisfied for other statistics. More research is needed to assess whether the BH procedure is robust when independence and positive regression dependency are violated.

What Are Problems with These Solutions?

There are two related concerns with the adjustment procedures discussed above: (1) they result in tests with reduced statistical power and (2) they could result in tests with even less power when the test statistics are correlated (dependent).

Losses in Statistical Power

The statistical procedures that control for multiplicity reduce Type I error rates for individual tests. Consequently, these adjustment procedures result in tests with *reduced* statistical power—the probability of rejecting the null hypothesis given that the null hypothesis is false. Stated differently, these adjustment methods reduce the likelihood that the tests will identify *true* differences between the contrasted groups. The more conservative the multiple testing strategy, the greater the power loss.

Table B.4 demonstrates power losses based on the simulations discussed above when the FWER is controlled using the Bonferroni and Holm procedures and the FDR is controlled using the BH procedure.

Table B.4: Statistical Power with Multiplicity Adjustments

		Statistical Power			
			FWER Control		FDR Control
Number of Tests	Percentage of Tests with True Null Hypotheses	*No* Adjustments	Bonferroni	Holm	Benjamini-Hochberg (BH)
5	80	0.80	0.59	0.59	0.55
10	80	0.80	0.50	0.50	0.55
20	80	0.80	0.41	0.42	0.55
50	80	0.80	0.31	0.32	0.55
5	50	0.80	0.59	0.61	0.67
10	50	0.80	0.50	0.53	0.67
20	50	0.80	0.41	0.44	0.67
50	50	0.80	0.31	0.33	0.67
5	20	0.80	0.59	0.66	0.74
10	20	0.80	0.50	0.57	0.74
20	20	0.80	0.41	0.47	0.74
50	20	0.80	0.31	0.35	0.74

Notes: Error rates and power levels were calculated using simulated data as described in the text. The calculations assume independent test statistics.

The key findings from the table are as follows:

- ***Power losses can be large using the Bonferroni and Holm procedures.*** The power of each method decreases with the number of tests. In the absence of multiple comparison adjustments, statistical power is 80 percent. Applying the Bonferroni correction reduces the power to 59 percent for 5 tests, 41 percent for 20 tests, and 31 percent for 50 tests. The Holm and Bonferroni procedures yield tests with similar power, but the Holm procedure performs slightly better if many impacts truly exist.

- ***The power of the BH procedure increases with the number of intervention effects that truly exist.*** Statistical power is 55 percent if 80 percent of null hypotheses are true, compared to 74 percent if only 20 percent of null hypotheses are true. The power of the BH procedure does not vary with the number of tests.

- ***Power losses are smaller for the BH procedure than for the Bonferroni and Holm procedures.*** Differences between the procedures become larger as (1) the number of tests increases and (2) the number of true null hypotheses decreases. Thus, power losses can be considerably smaller under the BH procedure if many contrasts truly differ.

These results suggest that multiplicity adjustments involve a tradeoff between Type I and Type II error rates. Conservative testing strategies, such as the Bonferroni and similar methods, can result in considerable losses in the statistical power of the tests, even if only a small number of tests are performed. The less conservative BH test has noticeably more power if a high percentage of all null hypotheses are false.

Dependent Test Statistics

Individual test statistics are likely to be related in many evaluations of educational interventions. Consider testing for intervention effects across many outcomes measured for the same subjects. In this case, the test statistics are likely to be correlated, because a common latent factor may affect the outcomes for the same individual and treatment effects may be correlated across outcomes. As another example, if multiple treatment alternatives are compared to each other or to the same control group, the test statistics are correlated because of the overlap in the samples across pairwise contrasts.

Some of the adjustment methods discussed above (such as the Bonferroni and Holm methods) control the FWER at a given α level when tests are correlated. However, for some forms of dependency, these methods may adjust significance levels for individual tests by more than is necessary to control the FWER. This could lead to further reductions in the statistical power of the tests. For example, if test correlations are positive and large, each test statistic is providing similar information about intervention effects, and thus, would likely produce similar *p*-values. Consequently, in these situations, fewer adjustments to Type I error rates are needed to control the FWER.

This problem can be demonstrated using the simulations discussed above. FWER values were calculated for 10 tests, all with true null hypotheses, when the correlation between all test statistics (ρ) ranged from 0 to 1. FWER values were calculated for unadjusted *t*-tests and using the Bonferroni and Holm adjustment methods (Table B.5).

Table B.5: FWER Values for 10 Positively Correlated Tests

Correlation Between Tests (ρ)	FWER for Unadjusted Tests	Adjusted Tests	
		Bonferroni Method	Holm Method
0.0	0.40	0.05	0.05
0.2	0.38	0.05	0.05
0.4	0.32	0.04	0.04
0.6	0.24	0.03	0.03
0.8	0.18	0.02	0.02
1.0	0.05	0.005	0.005

Notes: Error rates were calculated using simulated data as described in the text. The calculations assume that all null hypotheses are true.

The unadjusted FWERs become smaller as ρ increases (Table B.5). The FWER reduces from 0.40 for independent tests ($\rho = 0$) to 0.24 when $\rho = 0.6$ to 0.18 when $\rho = 0.8$. As a result, the Bonferroni and Holm methods "overcorrect" for multiplicity in this example and yield test statistics with reduced power.

Several methods discussed above adjust for dependency across test statistics. For example, the resampling methods incorporate general forms of correlational structures across tests, and the Tukey-Kramer, Dunnett, and Orthogonal Contrast methods account for specific forms of dependency when various treatments are compared to each other or to a common control group. Thus, power gains can be achieved using these methods. In addition, as discussed, the BH method controls the FDR under certain forms of dependency and for certain test statistics, but not for others.

Summary and Recommendations

The testing guidelines discussed in this report minimize the extent to which multiple testing adjustment procedures are needed by focusing on significance tests for composite domain outcomes. However, adjustment procedures for individual tests are needed for some testing situations, such as between-domain analyses, subgroup analyses, and designs with multiple treatment groups. Thus, this section provides general recommendations on suitable methods, although it should be emphasized there is not one statistical procedure that is appropriate for all settings; applicable methods will depend on the key research questions and the structure of the hypothesis tests.

To control the FWER, the bootstrap or permutation resampling methods are applicable for many testing situations because they incorporate general distributional and dependency structures across tests (Westfall and Young 1993). In educational evaluations, correlated test statistics are likely to be common. Thus, the resampling methods are recommended because they could yield tests with greater statistical power than other general-purpose methods that typically do not adjust for dependent data. The main disadvantages of the resampling methods are that they are difficult to explain and are computer intensive.

For FWER control, the Holm (1979) and Bonferroni procedures may also be suitable general-purpose methods. These methods (and especially the Bonferroni procedure) are easier to explain and apply than the resampling methods. However, although the Bonferroni and Holm methods control the FWER for dependent test statistics, they *do not account* for the dependency structure across tests and, thus, tend to have lower statistical power than the resampling methods. Statistical power is somewhat greater for the Holm method than the Bonferroni method if many impacts truly exist across the contrasts.

Hsu (1996) recommends alternative procedures to control the FWER in certain testing situations. The Tukey-Kramer (Tukey 1953, Kramer 1956) method is recommended if *all* pairwise comparisons of means are of primary interest, and the Dunnett (1955) method is recommended if multiple treatments are compared to a common control group. These methods can yield greater statistical power than other methods because they account for the exact nature of the dependency across test statistics due to the repetition of samples across contrasts. The use of orthogonal contrasts is another possibility if they correspond to key research questions.

The BH procedure (Benjamini and Hochberg 1995) controls the FDR, which is a different criterion than the FWER for defining the overall Type I error rate across the family of tests. As discussed, if many beneficial impacts truly exist, the BH procedure tends to have more statistical power than the methods that control the FWER by allowing more Type I errors. The philosophy of the BH procedure is that if many impacts are found to be statistically significant, a few more false positives will not change the overall validity of study conclusions about intervention effects. However, if few impacts are statistically significant (signaling that many null hypotheses are true), the BH and FWER-controlling methods are similar.

There are several issues that need to be considered when using the BH procedure. First, it may not control the FDR for all forms of dependency across test statistics. Second, the BH method may be less appropriate for some confirmatory analyses than methods that control the FWER, because it applies a less stringent standard for controlling Type I errors. On the other hand, the BH method may be appealing because it operates under the philosophy that conclusions regarding intervention effects are to be based on the preponderance of evidence and could lead to increases in the statistical power of the tests. The choice of whether the study aims to control the FDR or FWER is an important design issue that needs to be specified prior to the data analysis.

Appendix C

Weighting Options for Constructing Composite Domain Outcomes

The testing guidelines discussed in this report focus on significance tests for composite domain outcomes, which are combinations of individual domain outcomes. Thus, a critical issue is how to weight individual domain outcomes to construct composites.

There is a large literature over many decades and across multiple disciplines on methods to combine multiple pieces of information to create composites (see, for example, Kane and Case 2004, Wainer and Thissen 1993, Wang and Stanley 1970, Gulliksen 1950, Wilks 1938). Similar to the multiple comparisons literature, there is no consensus on the "optimal" method that should be used to form composites that fits all circumstances. Rather, procedures should be selected that are best suited to the types of domain outcome measures that are under investigation and key research questions.

Composite formation rules should be specified in the study protocols. In developing rules, potential correlations among the domain outcomes should be considered. As discussed, outcome measures will likely be grouped into a domain if they are expected to measure a common latent construct. In situations where this objective is satisfied, different composite formation methods should yield similar composites (Landis et al. 2000). However, if the domain outcomes tap multiple factors, the choice of method could affect the resulting composites, in which case it may be appropriate to reconsider domain definitions. Thus, issues pertaining to the selection of weights for composite formation are similar to issues pertaining to the delineation of outcome domains.

This appendix briefly discusses composite formation methods that are found in the literature that fit our context. A full literature review is beyond the scope of this introductory discussion.

For the ensuing discussion, it is assumed that a domain contains N outcome measures for each sample member, and that the vector Y_i pertains to outcome measure values for outcome i. The outcomes are assumed to be standardized to have mean 0 and standard deviation 1 to avoid the composite being dominated by component outcomes with large variances, although the weighting schemes discussed below apply also to raw outcomes. A *composite* domain outcome, C, is defined as follows: $C = \sum_{i=1}^{N} w_i Y_i$, where w_i are "nominal" weights assigned to each outcome. The many interrelated methods of weighting that have been used in education research involve selecting the w_i s to maximize various criterion functions, as discussed next.

Regression weights. Suppose a pertinent outside data source contains information on a well-established observable validity criterion as well as the outcome measures under investigation. These data could then be used to construct weights based on the relationship between the criterion measure and the outcomes. Examples of external criteria are measures of school readiness, longer-term test scores, high school graduation status, college attendance status, and earnings.

Regression weights can be obtained by regressing the criterion measure (the dependent variable) on the component outcomes (the independent variables) using standard multivariate regression methods. The parameter estimates on the outcomes could then be used as weights for composite formation using the evaluation sample. The larger the correlation between an outcome and the criterion and the more

independent an outcome is of other outcomes, the larger is the weight, all else equal. This approach yields weights that minimize the mean squared prediction error in the estimation sample.

The advantage of this method is that predictive validity is often recognized as a more important criterion than reliability in evaluating measurement procedures (Kane and Case 2004, Wang and Stanley 1970). However, this method can be used only if pertinent data are available to estimate regression weights that apply to the population under investigation and that are based on large samples to ensure the stability of results. Furthermore, even if these conditions are met, the results need to be interpreted carefully, because the estimation sample used to develop the weights may not necessarily yield optimal weights for the population from which the sample is drawn or for other samples drawn from the same population (Raju et al. 1997).

Finally, this approach is likely to be most useful when the outcomes (predictors) are relatively independent. This independence condition, however, will not typically be satisfied in our context. Thus, the regression approach may be more suited to forming composites for a between-domain analysis than a within-domain analysis.

Natural or unit weights. For this approach, the *N* outcomes are simply summed or averaged to form the composite—that is, w_i is set to 1 (or a constant) for each outcome. This method is based on the "agnostic" criterion that each outcome is equally important. It has the advantage that is easy to apply and understand. Bobko et al. (2007) show that this approach can be appropriate under many circumstances.

The use of unit weights does *not* necessarily imply, however, that each outcome contributes equally to the overall variance of the composite. The contribution of Y_i to the variance of C is $(w_i^2 + \sum_{j \neq i}^{N} w_i w_j \rho_{ij})$, where ρ_{ij} is the correlation between Y_i and Y_j. With unit weights, this contribution reduces to $(1 + \sum_{j \neq i}^{N} \rho_{ij})$, so that the "effective" weight for each component outcome will depend on its average correlation with other component outcomes. If average correlations are similar across outcomes, the effective and nominal unit weights will be similar.

A variant of this method is to select weights to ensure that each variable contributes equally to the total variance of the composite. This can be done by setting *N* equations of the form $(w_i^2 + \sum_{j \neq i}^{N} w_i w_j \rho_{ij})$ to a constant and solving iteratively for w_i (Wilks 1938).

Expert judgment or subjective weights. Another approach to developing composites is to employ a content-oriented strategy in which outcomes are assigned to composites based on existing theory or rational judgment. Under this approach, theory or expert guidance obtained prior to data analysis is used to determine the relative "importance" of each outcome to the underlying domain construct. This approach could also be used if some outcomes have more "information" than others. For example, for combining tests, weights could be assigned based on the length of the tests or the nature of the questions (Wainer and Thissen 1993). For instance, larger weights could be assigned to multiple-choice than true-false questions.

Maximum reliability weights. Another approach is to select weights to maximize the reliability of the composite. This approach is often discussed in the test theory literature for combining achievement test scores or items (Kane and Case 2004, Wainer and Thissen 1993). This approach has received attention

because reliability of a measure is a necessary, although not sufficient, condition for validity of a measure.

Reliability is defined as the proportion of the total variance in the composite which is true-composite variance. Thus, maximum reliability weights can be found by maximizing the variance of the composite between subjects (between-subject variance) relative to the variance across outcomes within subjects (within-subject variance). Wang and Stanley (1970) and Gulliksen (1950) discuss procedures for obtaining these weights. Item response theory (IRT) is a more recent version of reliability weighting that simultaneously provides weights and a scale for the item responses (see, for example, Lord 1980).

Equal correlation weights. Another criterion function is to select weights that equalize the correlation between each outcome measure and the composite. The correlation between Y_i and the composite C can be expressed as follows:

$$(1) \quad \rho_{iC} = \frac{w_i + \sum_{j \neq i}^{N} w_j \rho_{ij}}{\sqrt{Var(C)}}.$$

Because the denominator in (1) is the same for each outcome, equal correlation weights can be calculated by setting N equations of the form $(w_i + \sum_{j \neq i}^{N} w_j \rho_{ij})$ to a constant and solving iteratively for w_i. This procedure is logically consistent only if all outcomes are positively correlated.

Factor analysis weights. Another approach for forming composites is to conduct a factor analysis on the component outcomes and to use the single factor solution as the composite outcome. If the data support a multi-factor solution, domain reconfigurations may be considered. Criteria for assessing the appropriate factor structure must be specified in the study protocols and adhered to in the analysis.

Alternatively, factor loadings from factor analyses conducted on other relevant datasets (perhaps with larger norming samples) could be applied as weights to form composites. Issues pertaining to the feasibility of this approach and the interpretation of results are similar to those discussed above for regression weighting.

Multivariate analysis of variance (MANOVA) weights. MANOVA methods are commonly used to control the Type I error rate when examining treatment effects on multiple outcome measures (see, for example, Harris 1975). In our context, the MANOVA approach would involve conducting omnibus *F*-tests to address the research question: Did the intervention have a statistically significant effect on *any* outcome measure within the domain? This is a different question than the one addressed by the composite *t*-test approach: Did the intervention have a statistically significant effect on a common domain latent construct?

Because they address different research questions, these two approaches lead to different weighting schemes for combining the outcome measures. Under the MANOVA approach, weights are found to maximize the *test statistics pertaining to impacts*, whereas under the composite *t*-test approach, weights are found to best identify a common domain construct.

To demonstrate the implied weighting scheme for the MANOVA approach, assume that the N domain outcomes for each subject are sampled from a joint multivariate normal distribution with mean vector μ_T

for m_T treatments and μ_C for m_C controls and common variance-covariance matrix Ω. Consider the composite *impact* estimate, *CI*:

$$CI = \sum_{i=1}^{N} w_i I_i = w'I ,$$

where I_i is the impact estimate (mean treatment-control difference) for outcome *i* and *I* and *w* are *N*x1 column vectors of impacts and weights, respectively. The squared *t*-statistic can then be written as follows:

$$(2)\quad t^2(w) = [\frac{m_T m_C}{(m_T + m_C)}](w'I)'[w'\hat{\Omega}w]^{-1}(w'I) ,$$

where $\hat{\Omega}$ is the usual estimator for Ω based on sample variances and covariances.

The omnibus *F*-statistic that is typically produced by statistical software packages can be obtained by first finding the weights w^* that *maximize* (2) subject to the normalizing restriction $w'^*\hat{\Omega}w^* = 1$, and then inserting w^* into (2). This procedure yields Hotelling's T^2 statistic:

$$T^2 = [\frac{m_T m_C}{(m_T + m_C)}]I'\hat{\Omega}^{-1}I ,$$

which is just a multiple of the usual *F*-statistic.[5]

Thus, all else equal, MANOVA methods tend to place more weight on standardized outcomes with larger impacts (positive or negative) than smaller impacts. Stated differently, this method tends to select weights to maximize the chance of finding significant impact findings.

The MANOVA approach is not recommended for the confirmatory analysis for several reasons. First, because domain outcomes are likely to tap the same underlying construct, it seems more appropriate to examine treatment effects on a composite measure of this construct than to test whether treatment effects exist for any of its components. A second reason is that it is difficult to develop a confirmatory theory that would result in outcomes being weighted according to the size of their impacts. Instead, the MANOVA procedure is a post hoc, data-driven method that is more suited to exploratory analyses.

[5] Specifically, $F = kT^2$ where $k = (m_T + m_C - N - 1)/N(m_T + m_C - 2)$; *F* is distributed as $F(N, m_T + m_C - N - 1)$.

Appendix D

The Bayesian Hypothesis Testing Framework

This appendix summarizes key features of the Bayesian testing approach, which is the main alternative to the classical testing approach. The use of these methods in IES studies is an important area for future research. Spiegelhalter et al. (1994) and Gelman and Tuerlinckx (2000) provide a more detailed discussion of the Bayesian framework.

In the Bayesian view, assessing the effects of an intervention is a dynamic process in which any individual study takes place in a context of continuously increasing knowledge. Initial beliefs about treatment effects are incorporated into the analysis and are expressed as a *prior* distribution. The prior distribution could be based on objective evidence or subjective judgment, and the shape and location of this distribution reflects the level of confidence in the prior information.

Using Bayes theorem, the prior distribution of the impact, $f(\delta_j)$, is combined with the conditional distribution of the observed data given the impact, $g(data \mid \delta_j)$, to obtain a *posterior* (updated) distribution of the treatment effect:

$$h(\delta_j \mid data) \propto g(data \mid \delta_j) f(\delta_j).$$

The Bayesian impact estimate is the mean of the posterior distribution. If both the prior and conditional distributions are normally distributed, the mean of the posterior distribution is a weighted average of the observed impact and the mean of the prior distribution (where weights are inversely related to variances of the likelihood and prior distributions). Thus, the Bayesian approach "shrinks" the observed impact estimate to the mean of the prior distribution. The Bayesian approach addresses the following question: "What is the updated evidence on the impact, once we combine the previous with the new evidence?"

Differences between the Bayesian and classical analyses include the incorporation of prior beliefs, the absence of *p*-values, and the absence of the idea of hypothetical repetitions of the sampling process. The posterior estimate of the impact and its uncertainty as measured by a *credibility interval* is analogous to the classical differences-in-means point estimate and its associated confidence interval. This credibility interval, however, has a direct interpretation in terms of belief; probabilistic statements can be made about the size of the impact. Those who *misinterpret* classical confidence intervals as the region in which the effect is likely to lie are, in essence, adopting a Bayesian point of view.

Gelman et al. (2007) argue that multiple comparisons issues typically are less of a concern in Bayesian modeling than in classical inference. This is because under the Bayesian approach, the impact estimates for the various contrasts and their credibility intervals are shifted toward each other. This leads to *wider* confidence bands under the Bayesian approach. For example, under the classical approach, the usual 95 percent confidence interval for the impact, $\delta_j = \mu_{Tj} - \mu_{Cj}$, is $[(\bar{y}_{Tj} - \bar{y}_{Cj}) \pm 1.96\sqrt{2\sigma^2 / n}]$, where $\bar{y}_{Tj}$ and $\bar{y}_{Cj}$ are sample means of the outcome measure for treatments and controls, respectively; σ^2 is the variance of the outcome measure; and n is the treatment (control) group sample size. If the likelihood and prior distributions are normally distributed, the 95 percent Bayesian credibility interval based on the posterior distribution is $[(\bar{y}_{Tj} - \bar{y}_{Cj}) \pm 1.96\sqrt{(2\sigma^2 / n)(1 + \frac{\sigma^2 / n}{\tau^2})}]$, where τ^2 is the variance of the prior

distribution for δ_j. Thus, statistical significance is *less* likely to be found under the Bayesian than classical approach. Consequently, the Bayesian approach is conservative and appropriately accounts for multiple comparisons in many instances.

More research is needed about the applicability of the Bayesian approach in IES-funded experimental studies. In particular, a critical issue is whether credible prior distributions on intervention effects can be specified. This will depend on the credibility of the empirical evidence on the effects of similar interventions to the ones being tested. It will also depend on the extent to which theory can be used to structure the multidimensional data so that empirical Bayes methods can be used to formulate prior distributions from the data. For example, prior distributions for a specific site (or outcome) could be estimated using the combined impact estimates for similar sites (or outcomes) if there is a theoretical justification for these groupings.

References

Altman, D.G., K.F. Schulz, and D. Moher (2001). "The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration." *Annals of Internal Medicine,* 134, 663-694.

Bechhofer, R. and C. Dunnett (1982). "Multiple Comparisons for Orthogonal Contrasts: Examples and Tables." *Technometrics,* 24(3), 213-222.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A New and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B,* 57, 1289-1300.

Benjamini, Y. and D. Yekutieli (2001). "The Control of the False Discovery Rate in Multiple Testing Under Dependency." *The Annals of Statistics,* 29(4), 1165-1188.

Bobko, P., P. Roth, and M. Buster (2007). "The Usefulness of Unit Weights in Creating Composite Scores." *Organizational Research Methods,* 109(4), 689-709.

Brookes, S.T., E. Whitley, T.J. Peters, P.A. Mulheran, M. Egger, and G. Smith (2001). "Subgroup Analyses in Randomized Controlled Trials: Quantifying the Risks of False-Positives and False-Negatives." *Health Technology Assessment,* 5(33), 1-49.

Committee for Proprietary Medicinal Products (CPMP) (2002). "Points to Consider on Multiplicity Issues in Clinical Trials." London: The European Agency for the Evaluation of Medicinal Products (EMEA).

Cook, R. and V. Farewell (1996). "Multiplicity Considerations in the Design and Analysis of Clinical Trials." *Journal of the Royal Statistical Society, Series A,* 159, 93-110.

Curran-Everett, D. (2000). "Multiplicity Comparisons: Philosophies and Illustration." *American Journal of Physiology,* vol. R1-R8.

Duncan, D.B. (1955). "Multiple Range and Multiple *F*-Tests." *Biometrics,* 11, 1-42.

Dunnett, C.W. (1955). "A Multiple Comparison Procedure for Comparing Several Treatments with a Control." *Journal of the American Statistical Association,* 50, 1096-1121.

Fisher, R.A. (1935). *The Design of Experiments.* Edinburgh and London: Oliver and Boyd.

Freeman, M.F. and J.W. Tukey (1950). "Transformations Related to the Angular and Square Root." *Annals of Mathematical Statistics,* 21, 607-611.

Gelman, A., J. Hill, and M. Yajima (2007). "Why We (Usually) Don't Worry About Multiple Comparisons." Columbia University Working Paper. New York: Columbia University.

Gelman, A. and H. Stern (2006). "The Difference Between "Significant" and "Not Significant" Is Not Itself Statistically Significant." *The American Statistician,* 60(4), 328-331.

Gelman, A. and F. Tuerlinckx (2000). "Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures." Columbia University Working Paper. New York: Columbia University.

Gordon, A., G. Glazko, Z. Qiu, and A. Yakovlev (2007). "Control of the Mean Number of False Discoveries, Bonferroni and Stability of Multiple Testing." *The Annals of Applied Statistics,* 179-190.

Gulliksen, H. (1950). *Theory of Mental Health.* New York: Wiley.

Harris, R.J. (1975). *A Primer of Multivariate Statistics.* New York: Academic Press, Inc.

Hochberg, Y. (1988). "A Sharper Bonferroni Procedure for Multiple Tests of Significance." *Biometrika,* 75, 800-802.

Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics,* 6, 65-70.

Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods.* London: Chapman and Hall.

Kane, M. and S. Case (2004). "The Reliability and Validity of Weighted Composite Scores." *Applied Measurement in Education,* 17(3), 221-240.

Keuls, M. (1952). "The Use of the 'Studentized Range' in Connection with an Analysis of Variance." *Euphytica*, 1, 112-122.

Kirk, M. (1994). *Experimental Design: Procedures for the Behavioral Sciences.* Pacific Grove, CA: Brooks/Cole.

Kramer, C.Y. (1956). "Extension of the Multiple Range Test to Group Means with Unequal Numbers of Replications." *Biometrics,* 12, 307-310.

Landis, R., D. Beal, and P. Tesluk (2000). "A Comparison of Approaches to Forming Composite Measures in Structural Equation Models." *Organizational Research Methods,* 3(2), 186-207.

Lang, T. and M. Secic (2007). *How to Report Statistics in Medicine,* 2nd ed. Philadelphia: American College of Physicians.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New Jersey: Lawrence Erlbaum Associates, Inc.

Newman, D. (1939). "The Distribution of the Range in Samples From a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation." *Biometrika,* 35, 16-31.

Raju, N., R. Bilgic, J. Edwards, and P. Fleer (1997). "Methodology Review: Estimation of Population and Cross Validity and the Use of Equal Weights in Prediction." *Applied Psychological Measurement*, 21, 291-305.

Rom, D.M. (1990). "A Sequentially Rejective Test Procedure Based on a Modified Bonferroni Inequality." *Biometrika,* 77, 663-665.

Rothwell, P.M. (2005). "Subgroup Analyses in Randomized Controlled Trials: Importance, Indications, and Interpretation." *The Lancet,* 365, 176-186.

Saville, D.J. (1990). "Multiple Comparison Procedures: The Practical Solution." *The American Statistician,* 44, 174-180.

Savitz, D. and F. Olshan (1995). "Multiple Comparisons and Related Issues in the Interpretation of Epidemiologic Data." *American Journal of Epidemiology,* 142(9), 904-908.

Scheffé, H. (1959). *The Analysis of Variance.* New York: John Wiley & Sons.

Shaffer, J. (1995). "Multiple Hypothesis Testing." *Annual Review of Psychology,* 46, 561-584.

Šidák, Z. (1967). "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association,* 62, 626-633.

Spiegelhalter, D., L.S. Freedman, and M. Parmar (1994). "Bayesian Approaches to Randomized Trials." *Journal of the Royal Statistical Society, Series A*, 357-416.

Storey, J.D. (2002). "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society, Series B,* 64, 479-498.

Tukey, J.W. (1953). "The Problem of Multiple Comparisons." In *Mimeographed Notes.* Princeton, NJ: Princeton University.

Wainer, H. and D. Thissen (1993). "Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction." *Applied Measurement in Education,* 6(2), 103-118.

Wang, M. and J. Stanley (1970). "Differential Weighting: A Review of Methods and Empirical Studies." *Review of Educational Research*, 40, 663-705.

Westfall, P.H, Y. Lin, and S. Young (1990). "Resampling-Based Multiple Testing." In *Proceedings of the Fifteenth Annual SAS Users Group International.* Cary, NC: SAS Institute, Inc., 1359-1364.

Westfall, P.H., R. Tobias, D. Rom, R. Wolfinger, and Y. Hochberg (1999). *Multiple Comparisons and Multiple Tests Using SAS.* Cary, NC: SAS Institute, Inc.

Westfall, P.H. and R.D. Wolfinger (1997). "Multiple Tests with Discrete Distributions." *The American Statistician,* 51, 3-8.

Westfall, P.H. and S.S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment.* New York: John Wiley & Sons.

Wilks, S. (1938). "Weighting Systems for Linear Functions of Correlated Variables When There Is No Dependent Variable." *Psychometrika,* 3, 23-40.