Washington State
Institute for
Public Policy

110 Fifth Avenue Southeast, Suite 214 • PO Box 40999 • Olympia, WA 98504-0999 • (360) 586-2677 • FAX (360) 586-2793 • www.wsipp.wa.gov

September 2006

STUDY DESIGN: BENEFITS AND COSTS OF K–12 EDUCATIONAL PROGRAMS AND SERVICES

Public K–12 education accounts for about 41 percent of Washington State's general fund expenditures.[1] The 2006 Washington State Legislature called for a systematic examination of this investment, complementing other statewide efforts such as Washington Learns.[2] Specifically, the Legislature directed the Washington State Institute for Public Policy (Institute) to "begin the development of a repository of research and evaluations of the cost-benefits of various K–12 educational programs and services."[3]

This report provides background and describes the methodology we are using for this study. Results will be presented in a second report due March 1, 2007. As directed in legislation, the Institute will then issue annual updates incorporating the latest research findings from around the country.

The complete text of the legislative study direction is provided in Exhibit 1. For more information, contact: Annie Pennucci at (360) 586-3952 or email: pennuccia@wsipp.wa.gov.

Background

Nationwide, experimental or rigorous quasi-experimental research is increasingly used by local, state, and federal policymakers to identify and implement effective public services. The Washington State Legislature has, in recent years, directed the Institute to examine "evidence-based" programs related to prevention, early intervention, mental health, substance abuse treatment, and criminal justice for both children

Exhibit 1
Legislative Study Direction

The 2006 Washington State Legislature directed the Institute to conduct a cost-benefit analysis of K–12 programs in Engrossed Substitute Senate Bill 6386 §607(15):

> *$125,000 of the general fund--state appropriation for fiscal year 2007 is provided solely for the Washington state institute for public policy to begin the development of a repository of research and evaluations of the cost-benefits of various K-12 educational programs and services. The goal for the effort is to provide policymakers with additional information to aid in decision making. Further, the legislative intent for this effort is not to duplicate current studies, research, and evaluations but rather to augment those activities on an on-going basis. Therefore, to the extent appropriate, the institute shall utilize and incorporate information from the Washington learns study, the joint legislative audit and review committee, and other entities currently reviewing certain aspects of K-12 finance and programs. The institute shall provide the following: (a) By September 1, 2006, a detailed implementation plan for this project; (b) by March 1, 2007, a report with preliminary findings; and (c) annual updates each year thereafter.*

[1] K–12 expenditures made up 40.7 percent of the 2005-07 state general fund budget. State of Washington Legislative Evaluation and Accountability Program (LEAP) Committee. *2006 Legislative Budget Notes,* 481.

[2] The Governor-led Washington Learns committee is conducting an 18-month review of Washington's education system structure and funding, covering early learning, K–12, and higher education. The review includes an examination of evidence-based K–12 programs and policies. Their report is due November 2006. See: http://www.washingtonlearns.wa.gov/work/default.htm

[3] ESSB 6386 §607(15), Chapter 372, Laws of 2006.

and adults.[4] In these previous studies, the legislature also asked the Institute to estimate the costs and benefits of research-based approaches. For this current assignment, the Institute will build on these earlier analyses to develop a comprehensive economic analysis of the costs and benefits of various evidence-based K–12 educational programs, policies, and services.

In addition to recent developments in Washington State, attention to experimental research in K–12 education has been magnified considerably since the passage of the federal No Child Left Behind (NCLB) Act in 2001. The Act defines "scientifically based research" as the "application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs."[5] The NCLB text expresses a preference for random-assignment experiments, stating that experimental or quasi-experimental research designs—those that include a control or comparison group—constitute the basis for sufficiently rigorous research. The Institute's cost-benefit work is based on the same principles of scientific research.

The next section outlines the Institute's research design for this cost-benefit analysis of K–12 educational programs and services.

Research Design

When we conduct a cost-benefit analysis, we focus on two major tasks. First, we locate evaluations of programs, policies, and services and assess each study for its methodological rigor. Studies that meet the Institute's criteria for scientific rigor (described in Appendix A) are synthesized to estimate average effectiveness. Second, we estimate the cost of implementing each program and the associated monetary benefits that accrue from any statistically significant outcomes identified in the first step.

Thus, the first step asks "Based on scientific evidence, what works?" and the second step asks "What is the return on investment?"

For this cost-benefit analysis of K–12 policies, programs, and services, each of these tasks is described briefly below. For technical readers, we also include two appendices that describe the Institute's methodology in greater detail.

Task 1. Meta-analyze the research literature. The first task is to conduct a scientific review of the research literature on K–12 educational programs and services. The purpose is to identify programs and policies shown to work, and, just as important, approaches that have been rigorously evaluated and found not to work.

The research approach we are employing in this study is called a "systematic review" of the evidence. In a systematic review, the results of all rigorous evaluation studies are analyzed to determine if, on average, it can be stated scientifically that a program achieves a particular outcome. A systematic review can be contrasted with a "narrative" review of the literature where a writer selectively cites studies to tell a story about a topic, such as crime prevention. Both types of reviews have their place, but systematic reviews are generally regarded as more rigorous and, because they assess all available studies and employ statistical hypothesis tests, they have less potential for drawing biased or inaccurate conclusions. Systematic reviews are being used with increased frequency in medicine, education, criminal justice, and many other policy areas.[6]

We are currently in the process of identifying, collecting, and interpreting the research results of a variety of K–12 evaluation studies. Our focus is on *educational* programs and services. That is, we are studying approaches where the primary purpose is to improve students' academic outcomes. Thus, our initial review will not include school-based programs that focus on non-academic outcomes such as teenage pregnancy or drug and alcohol abuse. These are important

[4] See: (a) S. Aos, M. Miller, and E. Drake (2006). *Evidence-based adult corrections programs.* Olympia: Washington State Institute for Public Policy; (b) S. Aos, J. Mayfield, M. Miller, and W. Yen (2006). *Evidence-based treatment of alcohol, drug, and mental health disorders: Potential benefits, costs, and fiscal impacts for Washington State.* Olympia: Washington State Institute for Public Policy; (c) S. Aos, R. Lieb, J. Mayfield, M. Miller, and A. Pennucci (2004). *Benefits and costs of prevention and early intervention programs for youth.* Olympia: Washington State Institute for Public Policy; and (d) S. Aos, P. Phipps, R. Barnoski, and R. Lieb (2001). *The comparative costs and benefits of programs to reduce crime.* Olympia: Washington State Institute for Public Policy.

[5] NCLB Act 2001, Title IX Part A §37(A).

[6] An international effort aimed at organizing systematic reviews is the Campbell Collaborative—a non-profit organization that supports systematic reviews in the social, behavioral, and educational arenas. See: http://www.campbellcollaboration.org.

questions, but for this review, we are focused on the question: What works to improve academic outcomes?

The programs, services, and policies included in the Institute's initial literature review are listed in Exhibit 2. Over time, this list of topics can be expanded to meet Washington's policy information needs. The legislative direction to the Institute was to "begin the development of a repository of research and evaluations" on what works—the programs, services, and policies listed in Exhibit 2 is our initial list for the March 2007 report.

The specific types of outcomes that we will meta-analyze depends on the measures used in existing K–12 evaluation studies. Again, the focus is on programs, services, and policies that aim to improve students' academic outcomes. These outcome measures include, but are not limited to, the following:

- Standardized test scores;
- Course grades or grade point averages;
- Grade retention;
- Years in special education;
- High school graduation/dropping out; and
- Longer-range outcomes such as college attendance, college graduation, employment, and earnings.

Appendix A provides technical detail on the Institute's meta-analysis methodology.

Task 2. Estimate the monetary benefits and costs associated with each evidence-based program. In addition to identifying programs that do and do not work, the Legislature also directed the Institute to determine the costs and benefits of K–12 programs and services.

Exhibit 2

Initial List of K–12 Educational Programs, Policies, and Services Included in the Literature Review

✓ Alternative learning environments (such as distance learning and alternative schools)
✓ Block scheduling
✓ Career and technical education
✓ Charter schools
✓ Class size
✓ Comprehensive school reform
✓ Dropout prevention
✓ English language learner instruction
✓ Extended learning options (including summer school, before/after school and Saturday programs, and a longer school day)
✓ Full-day kindergarten
✓ Gifted/talented student programs
✓ Instructional curricula
✓ Instructional technology
✓ Mentoring (for both teachers and students)
✓ Preschool (updated information from 2004 WSIPP cost-benefit analysis of prevention/early intervention programs)
✓ School-based health care
✓ School-based mental health care
✓ School counseling
✓ School leadership
✓ School lunch/breakfast
✓ School size
✓ Special education
✓ Student assessment policies/practices
✓ Teacher aides
✓ Teacher compensation policies
✓ Teacher instructional coaches
✓ Teacher professional development
✓ Tutoring
✓ Vouchers
✓ Year-round school calendar

To do this, we have developed, and are continuing to refine, techniques to measure costs and benefits associated with the programs, policies, and services listed in Exhibit 2. We will identify programs whose benefits outweigh their costs and also estimate the costs and benefits of programs that do not break even. In systematically examining the state's investment in K–12 education, it is just as important to know which strategies do not produce positive returns for the taxpayer's dollar as it is to identify those with proven positive returns.

As in our previous cost-benefit analyses, we will estimate costs and benefits in two ways: first, we estimate the benefits that accrue directly to program participants (in this case, the students), and second, we estimate the benefits that accrue to non-participants. For example, a student who graduates from high school enjoys the benefit of greater earning potential compared with students who do not graduate. Non-participants benefit from the higher taxes paid based on those increased earnings. Additionally, evidence exists for a causal link between high school graduation and subsequent reduced crime.[7] Examining costs and benefits from both the participant and taxpayer perspectives provides a more comprehensive description of cost-effectiveness.

The methodology for estimating costs and benefits associated with educational outcomes of K–12 programs and services is described in Appendix B.

Appendix A: Meta-Analysis Techniques

This technical appendix describes the study coding criteria, procedures for calculating effect sizes, and adjustments to effect sizes that will be used in the Institute's meta-analysis of K–12 educational programs and services.

Meta-Analysis Coding Criteria

The following are key coding criteria for our meta-analysis of evaluations of K–12 educational programs and services.

1. **Study Search and Identification Procedures.** We are currently searching for all K–12 evaluation studies completed since 1970 that are written in English. We are using three primary sources: a) study lists in other reviews of the K–12 research literature; b) citations in individual evaluation studies; and c) research databases/search engines such as Google, Proquest, Ebsco, ERIC, and SAGE.
2. **Peer-Reviewed and Other Studies.** Many K–12 evaluation studies are published in peer-reviewed academic journals, while others are from government or other reports. It is important to include non-peer reviewed studies, because it has been suggested that peer-reviewed publications may be biased toward positive program effects. Therefore, our meta-analysis will include studies regardless of their source.
3. **Control and Comparison Group Studies.** We only include studies in our analysis if they have a control or comparison group. We do not include studies with a single-group, pre-post research design. We believe that it is only through rigorous comparison group studies that average treatment effects can be reliably estimated.[8]
4. **Random Assignment and Quasi-Experiments.** Random assignment studies are preferred for inclusion in our review, but we also include studies with non-randomly assigned groups if sufficient information is provided to demonstrate comparability between the treatment and comparison groups.
5. **Exclusion of Studies of Program Completers Only.** We do not include a study in our meta-analytic review if the treatment group is made up solely of program completers. We believe there are too many significant unobserved self-selection factors that distinguish a program completer from a

[7] L. Lochner and E. Moretti (2004) The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review* 94(1): 155-189.

[8] See: *Identifying and implementing education practices supported by rigorous evidence: A user friendly guide* (2003, December) Coalition for Evidence-Based Policy, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available at: http://www.evidencebasedpolicy.org/docs/Identifying_and_Implementing_Educational_Practices.pdf

program dropout, and these factors are likely to bias estimated treatment effects. We do, however, retain such a study if sufficient information is provided to allow us to reconstruct an intent-to-treat group that includes both completers and non-completers, or if the demonstrated rate of program non-completion is very small (e.g. under 10 percent).

6. **Enough information to Calculate an Effect Size.** Following the statistical procedures in Lipsey and Wilson,[9] a study must provide the necessary statistical information to calculate an effect size. If such information is not provided, the study cannot be included in our review.

7. **Mean Difference Effect Sizes.** For this study we are coding mean difference effect sizes following the procedures in Lipsey and Wilson (2001). For dichotomous measures, we use the arcsine transformation to approximate the mean difference effect size, again following Lipsey and Wilson.

8. **Unit of Analysis.** Our unit of analysis is an independent test of treatment at a particular site. Some studies report outcome evaluation information for multiple sites; we include each site as an independent observation if a unique comparison group is also used at each site.

9. **Multivariate Results Preferred.** Some studies present two types of analyses: raw outcomes that are not adjusted for covariates, such as family income and ethnicity; and those that are adjusted with multivariate statistical methods. In these situations, we code the multivariate outcomes.

10. **Dichotomous Measures Preferred Over Continuous Measures.** Some studies include two types of measures for the same outcome: a dichotomous (yes/no) outcome and a continuous (mean number) measure. In these situations, we code an effect size for the dichotomous measure, because in small sample studies, continuous measures of education outcomes can be unduly influenced by a small number of outliers, while dichotomous measures can avoid this problem. Of course, if a study only presents a continuous measure, then we will code that measure.

11. **Longest Follow-Up Times.** When a study presents outcomes with varying follow-up periods, we generally code the effect size for the longest follow-up period. Since our intention for this analysis is to compute the long-run benefits and costs of different programs, we are interested in the longest follow-up time presented in evaluations.

12. **Some Special Coding Rules for Effect Sizes.** Most studies that meet the criteria for inclusion in our review will have sufficient information to code exact mean difference effect sizes. Based on other meta-analytic reviews we have completed, however, we anticipate that some studies will report some, but not all, of the information required. The rules we follow for these situations are as follows:

    a. **Two-Tail P-Values.** Sometimes, studies only report p-values for significance testing of program outcomes. If the study reports a one-tail p-value, we will convert it to a two-tail test.

    b. **Declaration of Significance by Category.** Some studies report results of statistical significance tests in terms of categories of p-values, such as p<=.01, p<=.05, or “not significant at the p=.05 level.” We calculate effect sizes in these cases by using the highest p-value in the category; e.g., if a study reports significance at “p<=.05,” we calculate the effect size at p=.05. This is the most conservative strategy. If the study simply states a result was “not significant,” we compute the effect size assuming a p-value of .50 (i.e. p=.50).

Procedures for Calculating Effect Sizes

Effect sizes measure the degree to which a program has been shown to change an outcome for program participants relative to a comparison group. There are several methods used by meta-analysts to calculate effect sizes, as described in Lipsey and Wilson (2001). In this meta-analysis, we are using statistical procedures to calculate the *mean difference effect sizes* of programs. We are not using the odds-ratio effect size because many outcomes measured in K–12 program evaluations are continuously measured. Thus, the mean difference effect size is a natural choice.

Many of the outcomes we are analyzing, however, are measured as dichotomies. For these yes/no outcomes, Lipsey and Wilson (2001) show that the mean difference effect size calculation can be approximated using the arcsine transformation of the difference between proportions.[10]

(A1) $$ES_{m(p)} = 2 \times \arcsin\sqrt{P_e} - 2 \times \arcsin\sqrt{P_c}$$

In this formula, $ES_{m(p)}$ is the estimated effect size for the difference between proportions from the research information; P_e is the percentage of the population that had an outcome such as high school graduation rates for the experimental or treatment group; and P_c is the percentage of the population that graduated.

A second effect size calculation involves continuous data where the differences are in the means of an outcome. When an evaluation reports this type of

[9] M. Lipsey and D. Wilson (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

[10] Ibid., Table B10, formula (22).

information, we use the standard mean difference effect size statistic.[11]

(A2) $$ES_m = \frac{M_e - M_c}{\sqrt{\frac{SD_e^2 + SD_c^2}{2}}}$$

In this formula, ES_m is the estimated effect size for the difference between means from the research information; M_e is the mean number of an outcome for the experimental group; M_c is the mean number of an outcome for the control group; SD_e is the standard deviation of the mean number for the experimental group; and SD_c is the standard deviation of the mean number for the control group.

Often, research studies report the mean values needed to compute ES_m in (A2) but fail to report the standard deviations. Sometimes, however, the research will report information about statistical tests or confidence intervals that can then allow the pooled standard deviation to be estimated. These procedures are also described in Lipsey and Wilson (2001).

Adjusting Effect Sizes for Small Sample Sizes. Since some studies have very small sample sizes, we follow the recommendation of many meta-analysts and adjust for this. Small sample sizes have been shown to upwardly bias effect sizes, especially when samples are less than 20. Following Hedges,[12] Lipsey and Wilson (2001)[13] report the "Hedges correction factor," which we use to adjust all mean difference effect sizes (N is the total sample size of the combined treatment and comparison groups):

(A3) $$ES'_m = \left[1 - \frac{3}{4N-9}\right] \times \left[ES_m, or, ES_{m(p)}\right]$$

Computing Weighted Average Effect Sizes, Confidence Intervals, and Homogeneity Tests. Once effect sizes are calculated for each program effect, the individual measures are summed to produce a weighted average effect size for a program area. We calculate the inverse variance weight for each program effect, and these weights are used to compute the average. These calculations involve three steps. First, the standard error, SE_m of each mean effect size is computed with:[14]

(A4) $$SE_m = \sqrt{\frac{n_e + n_c}{n_e n_c} + \frac{(ES'_m)^2}{2(n_e + n_c)}}$$

In equation (A4), n_e and n_c are the number of participants in the experimental and control groups and ES'_m is from equation (A3).

Next, the inverse variance weight w_m is computed for each mean effect size with:[15]

(A5) $$w_m = \frac{1}{SE_m^2}$$

The weighted mean effect size for a group of studies in program area *i* is then computed with:[16]

(A6) $$\overline{ES} = \frac{\sum (w_{m_i} ES'_{m_i})}{\sum w_{m_i}}$$

Confidence intervals around this mean are then computed by first calculating the standard error of the mean with:[17]

(A7) $$SE_{\overline{ES}} = \sqrt{\frac{1}{\sum w_{m_i}}}$$

Next, the lower, ES_L, and upper limits, ES_U, of the confidence interval are computed with:[18]

(A8) $$\overline{ES_L} = \overline{ES} - z_{(1-\alpha)}(SE_{\overline{ES}})$$

(A9) $$\overline{ES_U} = \overline{ES} + z_{(1-\alpha)}(SE_{\overline{ES}})$$

In equations (A8) and (A9), $z_{(1-\alpha)}$ is the critical value for the *z*-distribution (1.96 for $\alpha = .05$).

The test for homogeneity, which provides a measure of the dispersion of the effect sizes around their mean, is given by:[19]

(A10) $$Q_i = (\sum w_i ES_i^2) - \frac{(\sum w_i ES_i)^2}{\sum w_i}$$

The Q-test is distributed as a chi-square with *k*-1 degrees of freedom (where *k* is the number of effect sizes).

Computing Random Effects Weighted Average Effect Sizes and Confidence Intervals. When the p-value on the Q-test indicates significance at values of p less than or equal to .05, a random effects model is performed to calculate the weighted average effect size. This is accomplished by first calculating the random effects variance component, *v*.[20]

[11] Ibid., Table B10, formula (1).
[12] L. Hedges (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 6: 107-128.
[13] Lipsey and Wilson, *Practical meta-analysis*, 49, formula 3.22.
[14] Ibid., 49, equation 3.23.
[15] Ibid., 49, equation 3.24.
[16] Ibid., 114.
[17] Ibid.
[18] Ibid.
[19] Ibid., 116.
[20] Ibid., 134.

(A11) $$v = \frac{Q_i - (k-1)}{\sum w_i - (\sum wsq_i / \sum w_i)}$$

This random variance factor is then added to the variance of each effect size and then all inverse variance weights are recomputed, as are the other meta-analytic test statistics.

Adjustments to Effect Sizes

In the Institute's meta-analytic reviews, we make three types of adjustments to effect sizes that we believe are necessary to better estimate the results each program is likely to actually achieve in real-world settings. We make adjustments for:

- Methodological quality;
- Relevance or quality of the outcome measure(s) used; and
- Degree to which the researcher(s) who conducted the study was invested in the program's design and implementation.

Methodological Quality. Not all research is of equal quality, and this, we believe, greatly influences the confidence that can be placed in study results. Some studies are well designed and implemented, and the results can be viewed as accurate representations of whether the program worked. Other studies are not designed as well and less confidence can be placed in reported outcomes. In particular, studies of inferior research design cannot completely control for sample selection bias or other unobserved threats to the validity of reported research results. This does not mean that results from these studies are of no value, but it does mean that less confidence can be placed in any cause-and-effect conclusions.

To account for differences in the quality of research designs, we use a 5-point scale to adjust the reported results. The scale is based closely on the 5-point scale developed by researchers at the University of Maryland.[21] On this 5-point scale, a rating of "5" reflects an evaluation in which the most confidence can be placed. As the evaluation ranking gets lower, less confidence can be placed in any reported differences (or lack of differences) between the program and comparison/control groups.

On the 5-point scale, as interpreted by the Institute, each study is rated with the following numerical ratings.

- A **"5"** is assigned to an evaluation with well-implemented random assignment of subjects to a treatment group and a control group that does not receive the treatment/program. A good random assignment study should also indicate how well the random assignment actually occurred by reporting values for pre-existing characteristics of the program and control groups.
- A **"4"** is assigned to a study that employs a rigorous quasi-experimental research design with a program and matched comparison group, controlling with statistical methods for self-selection bias that might otherwise influence outcomes. These quasi-experimental methods may include estimates made with an instrumental variables modeling approach, or a Heckman approach to modeling self-selection.[22] A level 4 study may also be used to "downgrade" an experimental random assignment design that had problems in implementation, perhaps with significant attrition rates.
- A **"3"** indicates a non-experimental evaluation where the program and comparison groups are reasonably well matched on pre-existing differences in key variables. There must be evidence presented in the evaluation that indicates few, if any, significant differences observed in these salient pre-existing variables. Alternatively, if an evaluation employs sound multivariate statistical techniques (e.g. logistic regression) to control for pre-existing differences, then a study with some differences in pre-existing variables can qualify as a level 3.
- A **"2"** involves a study with a program and matched comparison group where the two groups lack comparability on pre-existing variables and no attempt was made to control for these differences in the study.
- A **"1"** involves a study where no comparison group is utilized. Instead, the relationship between a program and an outcome, e.g., grade point average, is analyzed before and after the program.

We do not use the results from program evaluations rated as a "1" on this scale, because they do not include a comparison group, and we believe that there is no context to judge program effectiveness. We also regard evaluations with a rating of "2" as highly problematic and, as a result, we do not consider their findings in the meta-analysis. In this study, we only consider evaluations that rate at least a 3 on this scale.

[21] L. Sherman, D. Gottfredson, D. MacKenzie, J. Eck, P. Reuter, and S. Bushway (1998). *Preventing crime: What works, what doesn't, what's promising*. Prepared for the National Institute of Justice. Department of Criminology and Criminal Justice, University of Maryland. Chapter 2.

[22] For a discussion of these methods, see W. Rhodes, B. Pelissier, G. Gaes, W. Saylor, S. Camp, and S. Wallace (2001). Alternative solutions to the problem of selection bias in an analysis of federal residential drug treatment programs. *Evaluation Review* 25(3): 331-369.

An explicit adjustment factor is assigned to each effect size based on the Institute's judgment concerning research design quality. We believe this adjustment is critical and is the only practical way to combine the results of a high quality study with those of lesser design quality.

- A level 5 study carries a factor of 1.0 (that is, there is no discounting of the study's evaluation outcomes).
- In our previous meta-analytic studies of other areas such as criminal justice, a level 4 study carried a factor of .75 (effect sizes discounted by 25 percent). We will re-evaluate the magnitude of this adjustment for the K–12 literature.
- In our previous meta-analytic studies of other areas such as criminal justice, a level 3 study carried a factor of .50 (effect sizes discounted by 50 percent). We will re-evaluate the magnitude of this adjustment for the K–12 literature.
- We do not include level 2 and level 1 studies in our analyses.

These factors are subjective to a degree; they are based on our researchers' general impressions of their confidence in the predictive power of studies of different quality. We also rely on evidence of the degree and direction of selection bias in non-random assignment studies to establish these adjustments.[23]

The effect of the adjustment is to multiply the effect size for any study, ES'_m, in equation (A3), by the appropriate research design factor. For example, if a study has an effect size of -.20 and it is deemed a level 4 study, then the -.20 effect size would be multiplied by .75 to produce a -.15 adjusted effect size for use in the cost-benefit analysis.

Adjusting Effect Sizes for Relevance or Quality of Outcome Measures. Our focus in this analysis is whether K–12 educational programs and services improve academic outcomes. We prefer measures such as test scores or grades and avoid measures such as tardiness or self-esteem, since these may or may not be related to long-run academic outcomes.

In addition, we require that all studies have at least a six-month follow-up period. For those studies that have a follow-up period of less than 12 months but more than six months, and for those studies that only use weak measures, we reduce effects sizes by 25 percent. This adjustment multiplies the effect size for any study with a short follow-up period or weak measure by .75.

Adjusting Effect Sizes for Researcher Involvement in Program Design and Implementation. The purpose of the Institute's work is to identify programs that can make cost-beneficial improvements to Washington's actual K–12 service delivery system. There is some evidence that programs that are closely controlled by researchers or program developers have better results than those in "real world" settings.[24] Therefore, we make an adjustment to effect sizes ES_m to reflect this distinction. As a parameter for all studies deemed not to be "real world" trials, the Institute discounts ES'_m by .5, although this can be modified on a study-by-study basis.

Appendix B: Estimating Costs and Benefits of Educational Outcomes

This technical appendix describes how the Institute has previously estimated the costs and benefits associated with various educational outcomes.[25] These methods will be re-examined and refined for this study.

Valuation of Education Outcomes

Many K–12 educational outcomes are measures of human capital: graduation from high school, number of years of schooling completed, and achievement test scores earned during the K–12 years. Other often-measured educational outcomes relate to the use of certain K–12 resources: years of special education and grade retention. The benefits associated with each of these possible outcomes are discussed in this section. Exhibit 3 lists the equations involved in the procedures to calculate economic values for these outcomes.

Human Capital Outcomes. Our approach estimates the value of changes in high school graduation rates, years of education completed, and achievement test scores during the K–12 years by estimating the expected change in lifetime earnings caused by a change in the human capital measure. Measuring the earnings implications of these human capital variables is a commonly used approach in economics.[26]

[23] M. Lipsey (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The Annals of the American Academy of Political and Social Science* 587(1): 69-81. Lipsey found that, for juvenile delinquency evaluations, random assignment studies produced effect sizes only 56 percent as large as nonrandom assignment studies.

[24] Ibid. Lipsey found that, for juvenile delinquency evaluations, programs in routine practice (i.e., "real world" programs) produced effect sizes only 61 percent as large as research/demonstration projects. See also: A. Petrosino and H. Soydan (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology* 1(4): 435-450.

[25] Aos, et al. (2004). *Benefits and costs of prevention and early intervention programs for youth.*

[26] See, for example, A. Krueger (2003) Economic considerations and class size. *The Economic Journal* 113(485): F34-F63, accessed from the author's website: http://edpro.stanford.edu/eah/eah.htm; and E. Hanushek (2003, October) *Some Simple Analytics of School Quality*, accessed from the author's website: http://edpro.stanford.edu/Hanushek/files_det.asp?FileId=139

In this analysis, all human capital earnings estimates derive from a common dataset. The estimates are taken from the latest U.S. Census Bureau's March Supplement to the Current Population Survey, which provides cross-sectional data for earnings by age and by educational status.[27] To these data we apply different measures of the net advantage gained through increases in each human capital outcome.

For the human capital high school graduation outcome, the CPS money earnings data, by age, are differenced between those who graduate from high school (with no further degree), *Earnhsgrad*, and those with less than a high school diploma, *Earnnonhsgrad*. This differenced series is then present valued to age 18 by applying the general real discount rate used in the overall analysis, *Dis*, and any assumed real rate of growth in wages, *Earnesc*. We use age 65 as the cut-off point for earnings.

These earnings in equation (B1) are then present valued further to the age of the person in the program, *progage*. The values are also converted to the base year dollars chosen for the overall cost-benefit analysis, IPD_{base}, relative to the year in which the CPS data are denominated, IPD_{cps}. A fringe benefit rate is applied to the earnings, *Fringe*. As mentioned, the model can accommodate a rough estimate of any non-market (i.e., non-earnings) outcomes that may be causally related to education outcomes; this is modeled with the *NonMarket* parameter in equation (B2). Additionally, since the observed difference between the wages of these two groups may not be all due to the causal factor of earning a high school diploma, a multiplicative causation/correlation factor, *HSgradCC* (with a value greater than or equal to zero, or less than or equal to one), can be applied to the present value to provide an estimate of the causal effect.[28]

For the human capital achievement test score outcome, a similar process is used. The CPS money earnings data, by age, are taken as a weighted average of those with a high school diploma and those with some college, *Earnhsgradplus*, but not a college degree. This stream of earnings is multiplied by an estimated rate of return to earnings per one standard deviation increase in achievement test scores, *TestScoreROR*.[29] We calculate a present value to age 18 by applying the general real discount rate used in the overall analysis, *Dis*, and any assumed real rate of growth in wages, *Earnesc*. We use age 65 as the cut-off point for earnings. The remaining calculations in equation (B4) follow the procedures discussed for equation (B2).

For the human capital number of years of education outcome, the process is exactly the same. The CPS money earnings data, by age, are taken as a weighted average of those with a high school diploma and those with some college but no degree, *Earnhsgradplus*. This stream of earnings is multiplied by an estimated rate of return to earnings per extra year of formal education, *EdyearsROR*. The remaining calculations in (B6) follow those discussed for equation (B2).

Some K–12 programs, policies, and services we will evaluate include more than one of these human capital variables. For example, some K–12 evaluations produce effect sizes for high school graduation and for K–12 test scores. In these cases, we only include one of the human capital variables, and we use the outcome that produces the highest economic return.

K–12 Resource Outcomes. The model can also calculate the value of two other often measured K–12 educational outcomes: years of special education and grade retention. The present value costs of a year of special education is estimated by discounting the cost of a year in special education, *SpecEdCostYear*, for the estimated average number of years that special education is used, conditional on entering special education, *specedyears*. These years are assumed to be consecutive. The present value is to the age when special education is assumed to first be used, *start*. In equation (B8), this sum is further present valued to the age of the youth in a program, *progage*, and the cost is expressed in the dollars used for the overall cost benefit analysis, *IPDbase*, relative to the year in which the special education costs per year are denominated, *IPDspecedcostyear*.

The present value cost of an extra year of K–12 education is estimated for those retained for an extra year. This is modeled by assuming that the cost of the extra year of K–12 education, *EdCostYear*, after adjusting the dollars to be denominated in the base year dollars used in the overall analysis, would be borne when the youth is approximately 18 years old. Since there is a chance that the youth will not finish high school and, therefore, that the cost of this year will never be incurred, this present valued sum is multiplied by the probability of high school completion, *Hsgradprob*.

[27] The data are from the March Supplement to the CPS, PINC-04. Educational Attainment—People 18 Years Old and Over by Total Money Earnings, Age, Race, Hispanic Origin, and Sex.

[28] These types of causation/correlation adjustments have also been made in other cost-benefit analyses to avoid overstating benefits due to some unobserved selection bias. See, for example, M. Cohen (1998) The monetary value of saving a high risk youth. *Journal of Quantitative Criminology* 14(1): 5-33.

[29] Hanushek (2003). *Some Simple Analytics of School Quality*.

Other Outcomes Linked to Human Capital Outcomes. Research literature has also focused attention on several types of non-market benefits associated, perhaps causally, with the human capital outcomes evaluated in this analysis. A listing of possible non-market benefits to education appears in the work of Wolfe and Haveman.[30] In our current cost-benefit model, we do not estimate these non-earnings values explicitly, with one exception (discussed below). Rather, we provide a simple multiplicative parameter that can be applied to the estimated earnings effects so that the non-market benefits can be roughly modeled. Since some research indicates that these non-market benefits of human capital outcomes can be considerable, future refinements to our cost-benefit model will attempt to analyze these possible non-wage benefits explicitly.

The one exception that we model explicitly in this analysis is the relationship between high school graduation rates and their independent causal effect on crime. This conclusion is based on a recent study by Lochner and Moretti.[31] Their work offers convincing evidence of a statistically significant, albeit relatively weak, link between high school graduation and subsequent reduced crime. They use a variety of econometric methods and several nationally representative datasets to estimate this relationship. We calculated an effect size of the relationship from the Lochner and Moretti study to be -.061. To put that effect size in perspective, we found that some programs for juvenile offenders (i.e., programs that focus on higher-risk youth) can reduce subsequent crime with an effect size of -.188.

Exhibit 3 on the following two pages displays the equations used to calculate the present value costs of these education outcomes.

[30] B. Wolfe and R. Haveman (2002) “Social and nonmarket benefits from education in an advanced economy.” Proceedings from the Federal Reserve Bank of Boston's 47th economic conference, *Education in the 21st Century: Meeting the Challenges of a Changing World*, accessed from: http://www.bos.frb.org/economic/conf/conf47/index.htm. See also a collection of articles on the topic published in J. Behrman and N. Stacey, eds. (1997). *The social benefits of education*. Ann Arbor: The University of Michigan Press.

[31] L. Lochner and E. Moretti (2004). The effect of education on crime.

Exhibit 3

Equations Used to Calculate the Present Valued Costs of Education Outcomes

High School Graduation

(B1) $$PVEarn_{18}=\sum_{y=18}^{65}\frac{(Earnhsgrad_{y}-Earnnonhsgrad_{y})\times(1+Earnesc)^{y-17}}{(1+Dis)^{y-17}}$$

(B2) $$PVEarn_{progage}=\frac{PVEarn_{18}\times\frac{IPD_{base}}{IPD_{cps}}\times(1+Earnesc)^{18-progage}\times Fringe\times(1+NonMarket)\times HSgradCC}{(1+Dis)^{18-progage}}$$

Test Scores

(B3) $$PVEarn_{18}=\sum_{y=18}^{65}\frac{(Earnhsgradplus_{y}\times TestScoreROR_{y})\times(1+Earnesc)^{y-17}}{(1+Dis)^{y-17}}$$

(B4) $$PVEarn_{progage}=\frac{PVEarn_{18}\times\frac{IPD_{base}}{IPD_{cps}}\times(1+Earnesc)^{18-progage}\times Fringe\times(1+NonMarket)\times TestScoreCC}{(1+Dis)^{18-progage}}$$

Years of Education

(B5) $$PVEarn_{18}=\sum_{y=18}^{65}\frac{(Earnhsgradplus_{y}\times EdyearsROR_{y})\times(1+Earnesc)^{y-17}}{(1+Dis)^{y-17}}$$

(B6) $$PVEarn_{progage}=\frac{PVEarn_{18}\times\frac{IPD_{base}}{IPD_{cps}}\times(1+Earnesc)^{18-progage}\times Fringe\times(1+NonMarket)\times EdyearsCC}{(1+Dis)^{18-progage}}$$

Earnhsgrad = The annual CPS earnings of high school graduates. Annual money earnings of an individual in year y, taken from the U.S. Census Bureau’s March Current Population Survey, Annual Demographic Supplement, Table: PINC-04. Educational Attainment—People 18 Years Old and Over, by Total Money Earnings, Age, Race, Hispanic Origin, and Sex.

Earnnonhsgrad = The annual CPS earnings of non high school graduates, same source as above.

Earnhsgradplus = The annual CPS earnings of high school graduates plus those with some college but no degree, same source as above.

Earnesc = An estimated long-run annual growth rate in real earnings.

IPD_{base}, IPD_{cps} = The implicit price deflator for the year chosen as the base year for the overall analysis and for the year in which the current population survey is based.

NonMarket = An estimate of the non-earnings benefits of education expressed as a percentage of the earnings effect.

HSgradCC = A causation-correlation factor for high school graduation to adjust the cross-sectional CPS data.

TestScoreCC = A causation-correlation factor for test scores to adjust the cross-sectional CPS data.

EdyearsCC = A causation-correlation factor for the number of years of education to adjust the cross-sectional CPS data.

Fringe = The fringe benefit rate used in the analysis.

Taxrate = The tax rate used in the analysis.

TestScoreROR = The annual rate of return for a one standard deviation increase in achievement test scores.

EdyearsROR = The annual rate of return for an extra year of education.

progage = The average age of a youth in a program.

Dis = The real discount rate.

Exhibit 3, Continued

Equations Used to Calculate the Present Valued Costs of Education Outcomes

K–12 Special Education

(B7) $$PVspeced_{start} = \sum_{y=1}^{specedyears} \frac{SpecEdCostYear}{(1+Dis)^{y}}$$

(B8) $$PVspeced_{progage} = \frac{PVspeced_{start} \times \dfrac{IPD_{base}}{IPD_{speced\cos tyear}}}{(1+Dis)^{start-progage}}$$

K–12 Grade Retention

(B9) $$PVgraderet_{progage} = \left[\frac{EdCostYear \times \dfrac{IPD_{base}}{IPD_{ed\cos tyear}}}{(1+Dis)^{18-progage}} \right] \times Hsgradprob$$

SpecEdCostYear = The incremental cost of a year of special education, compared to a year of regular K–12 education.

specedyears = The average number of years in special education for a youth who enters special education.

start = The average age of a youth who starts special education.

progage = The average age of a youth in a program.

IPD_{base}, IPD = The implicit price deflator for the year chosen as the base year for the overall analysis, and for other costs.

EdCostYear = The cost of a year of regular K–12 education.

Hsgradprob = The probability that a youth who is retained sometime during the K–12 years will still be in school during his or her senior year in high school.

Dis = The real discount rate.

Document No. 06-09-2201

Washington State
Institute for Public Policy

The Washington Legislature created the Washington State Institute for Public Policy in 1983. A Board of Directors—representing the legislature, the governor, and public universities—governs the Institute and guides the development of all activities. The Institute's mission is to carry out practical research, at legislative direction, on issues of importance to Washington State.