Block Scheduling in the High School Setting

A Synthesis of Evidence-Based Research

Chance W. Lewis
Marc A. Winokur
R. Brian Cobb
Colorado State University

Gail S. Gliner
Metropolitan State College of Denver

Joel Schmidt
Ludwig-Maximilians-Universitat

February 2005

A report prepared for
MPR Associates, Inc., Berkeley, CA
under contract to the Office of Vocational and Adult Education,
U.S. Department of Education

Foreword

In 2003, in association with and under contract to the Office of Vocational and Adult Education within the U.S. Department of Education, MPR Associates, Inc. commissioned four papers on topics related to improving secondary education and student achievement. The topics included block scheduling, smaller learning communities, remediation or assistance programs aimed at ninth-graders, and school choice.

The authors of each paper applied a recently developed review and synthesis tool proffered by the What Works Clearinghouse, established by the Institute of Education Sciences. The What Works Clearinghouse gathers studies of the effectiveness of educational interventions, reviews the studies that have the strongest designs, and reports on the strengths and weaknesses of those studies against a specific set of Evidence Standards.

The resulting set of four research syntheses documents the degree to which each area of study includes research that achieves the level of rigor required to meet the standards, and whether the available research provides the clear evidentiary foundation necessary for drawing conclusions about each intervention's efficacy.

A subtask within contract ED-99-CO-0160 (Richard Smith, OVAE project director) funded the development of these papers. Opinions expressed and conclusions drawn in the research syntheses do not represent official U.S. Department of Education position or policy, nor of MPR Associates, Inc.

Table of Contents

Page

Introduction

The purpose of this study was to produce a systematic review and synthesis of evidence-based research on the effect of block scheduling on student achievement in United States high schools. This report provides a brief introduction to block scheduling, chronicles the search strategies used to locate the final literature set, and describes the processes employed to code the studies on outcome, intervention, and methodological criteria using the What Works Clearinghouse (WWC) framework. In addition, findings, conclusions, and recommendations are discussed for the studies that merited inclusion into the block scheduling evidence base.

Topic Description

Block scheduling first appeared in the literature over thirty years ago as modular scheduling, flexible scheduling, or modular flexible scheduling (Stewart & Shank, 1971). It was not until the late 1980s that it became one of the fastest growing educational reforms in U.S. secondary public schools. As of 1994, Cawelti estimated that almost 40 percent of American high schools had implemented some form of block scheduling. According to Rettig and Canady (2001), this trend continues unabated into the new millennium with 75 percent of high schools in states such as Virginia using block scheduling. Although a block scheduling intervention can be implemented in many different ways, all variations have the commonality of increasing the time available for instruction by extending classes beyond the traditional 50 minutes (Weller & McLeskey, 2000).

The most popular manifestation of block scheduling is the 4X4 Semester plan, also known as "accelerated schedule" or "Copernican." In a 4X4 block schedule, students can complete four yearlong equivalent courses in one semester by attending the same four 90-minute classes every day of the week. However, the amount of actual class time in a 4X4 course may be slightly less than in a traditionally scheduled course (Queen, Algozzine, & Eaddy, 1997). Another common type of block scheduling is the Alternate Day plan, also known as A/B, Odd/Even, or Day 1/Day 2. With A/B block scheduling, students take four 90-minute classes on alternating days for an entire school year. Although the amount of instructional time is comparable to 4X4 block scheduling, an Alternate Day plan gives students the full 180-day school year to complete the same eight courses (Brake, 2000). In addition, there are hybrid

block scheduling plans that combine elements of 4X4, A/B, and traditional scheduling formats.

For this systematic review, all variations of block scheduling were examined as a single intervention. The rationale for this decision was twofold. First, the extension of class time is the "essence" of every block scheduling format and is the criterion by which most of the studies in the evidence base defined the treatment in their analyses (e.g., Rice, Croninger, & Roellke, 2002). Second, with only seven studies in the evidence base, a more fine-grained analysis based on the type of block scheduling intervention would yield fewer studies for each alternative. Although the primary effect size analysis required that all block scheduling groups be combined into one treatment, a sensitivity analysis was conducted that allowed for comparisons between each block scheduling type and traditional scheduling.

Major Outcomes

Most of the published research within the past ten years is grounded in teacher and student perceptions of block scheduling culled from surveys and interviews (Nichols, 2000). The major outcomes from these descriptive studies are opinions on classroom climate (e.g., student-teacher relationships, discipline problems), instructional approaches, student development, and satisfaction with block scheduling. Other featured outcomes include beliefs about course scheduling and availability (e.g., foreign languages), student attention and retention spans, and the amount and type of content covered in block scheduling.

While qualitative data on the efficacy of block scheduling is plentiful, there is a dearth of experimental research on the academic achievement of students in block scheduling plans (O'Neil, 1995; Wallinger, 2000). Although test scores are the outcome of interest for this evidence report, the majority of quantitative data on block scheduling is based on student grades and attendance, graduation, retention, and discipline rates.

Literature Summary

The following literature summary reviews block scheduling research not meant to be included in a WWC Evidence Report. For example, studies based on interview and observational data, as well as empirical studies with quantitative outcomes other than standardized test scores are summarized. Grouping these studies by methodology allows for a closer examination of the objectives, outcomes, and findings of the block scheduling research base.

The majority of literature on block scheduling is in the form of survey research, evaluation reports, and qualitative case studies (Stanley & Gifford, 1998; Wronkovich, 1998). The pri-

mary emphasis of non-experimental research on block scheduling is stakeholder satisfaction. Most researchers have found that principals, teachers, and students are very satisfied with block scheduling (e.g., Hamdy & Urich, 1998). Furthermore, the amount of satisfaction seems to increase the longer that block scheduling is used (Edwards, 1995; Staunton, 1997). As for classroom climate, teachers perceive student/teacher relationships to be improved because there is more time for concentrated interactions (Eineder & Bishop, 1997; Skrobarcek et al., 1997). Parents have contributed to research on this topic by consistently reporting growth in the academic and social development of students participating in a block scheduling plan (Eineder & Bishop, 1997; Thomas & O'Connell, 1997).

The relationship between block scheduling and classroom instruction is another popular outcome for non-experimental research. Many studies have found that teachers appreciate the flexibility of block scheduling in providing longer planning periods, greater course offerings, and more time for in-depth study (e.g., Queen et al., 1997). For example, teachers have the time and support to develop curricula focused on cooperative learning exercises (Weller & McLeskey, 2000) and student-directed activities (Shortt & Thayer, 1995). Furthermore, teachers are encouraged to experiment with the pace of instruction to take advantage of the longer blocks of time afforded by the 4X4 and A/B plans (Pisapia & Westfall, 1997). Finally, Wilson and Stokes (2000) argue that students perceive block scheduling to be an effective approach if teachers use a greater variety of instructional strategies in the classroom.

The literature also contains some negative findings related to block scheduling interventions. For example, students identify the difficulty in making up work and the amount of busy work as the main disadvantages of block scheduling (Wilson & Stokes, 2000). Furthermore, students often report being more tired and less attentive during the longer class periods (Lapkin, Harley, & Hart, 1997). One of the most frequently perceived problems with the 4X4 Semester plan is the time gap between courses, in that a 4X4 class offered in the fall of one year may not be followed up until the spring semester of the next school year (Hamdy & Urich, 1998). Shortt and Thayer (1995) also found that teachers believe students need daily instruction in a subject to maximize their learning.

The findings from experimental and quasi-experimental studies have generally been positive for the effect of block scheduling on student grades, attendance rates, and graduation rates. Most researchers have reported statistically significant grade-point average (GPA) gains for students on a block schedule (e.g., Deuel, 1999; Edwards, 1995), while only some have found no effects, or adverse effects, for block scheduling students (e.g., Skrobarcek et al., 1997). This pattern continues with the findings for attendance and graduation, as a majority of studies have shown significant increases in daily attendance and student graduation rates

after conversion to a 4X4 (Nichols, 2000; O'Neil, 1995) or A/B block scheduling plan (Buckman, King, & Ryan, 1995).

The research on student discipline is decidedly mixed. Some studies have shown significant drops in discipline problems (e.g., suspensions) with block scheduling (Buckman et al., 1995; Eineder & Bishop, 1997; Thomas & O'Connell, 1997), while an equal number have reported no change in the amount of discipline incidents as compared with traditional scheduling (Deuel, 1999; Knight, DeLeon, & Smith, 1999; Wilson & Stokes, 1999).

Research Syntheses

Two recent research syntheses were reviewed to supplement the literature summary presented in this report. Rettig and Canady (2001) confirm much of the previous extant literature with an optimistic perspective regarding the effect of block scheduling on student outcomes. Specifically, they point to the unwavering support by all educational stakeholders for block scheduling as evidenced by the longevity of most interventions. They also note empirical findings that favor block scheduling students in regard to grades, honor roll placements, graduation rates, attendance, and discipline referrals. However, Rettig and Canady (2001) do not view block scheduling as a panacea and, in fact, identify numerous challenges including a lack of implementation fidelity.

The second review of literature by Stanley and Gifford (1998) presents a slightly more nuanced perspective, in that both the advantages and disadvantages of block scheduling are thoroughly articulated. For example, the upside of 4X4 block scheduling is that it "promotes student achievement by allowing the attendance of additional classes during the four-year high school tenure, by allowing more engaging learning activities, and by allowing students to concentrate narrowly on the four subjects taken each semester" (p. 10). However, Stanley and Gifford argue that this deeper approach to learning has a downside, in that teachers are unable to cover the breadth of content possible in a traditional schedule. This finding is the touchstone for one of the main controversies in the block scheduling field.

Controversies

There are several areas of controversy that have been identified through both quantitative and qualitative research on the topic. The first contentious issue is whether a decrease in quantitative minutes of classroom instruction is offset by the quality of student/teacher interactions in a block scheduling format (Nichols, 2000). Opponents argue that block sched-

uling is less effective because the difficulty in designing instruction appropriate for longer class periods inevitably results in the coverage of less material (O'Neil, 1995).

A second controversial topic is the timing of high-stakes tests relative to the completion of a 4X4 Semester block schedule. Specifically, there are concerns about the preparation of students who complete a 4X4 class in the fall and take Advanced Placement (AP) exams in the spring (O'Neill, 1995). In addition to questions about the scheduling of AP courses, the sequencing of foreign language and music classes is a topic of concern for the block scheduling intervention (Shortt & Thayer, 1995).

According to Rettig and Canady (2001), "the major remaining controversy still surrounding block scheduling is whether or not it will assist schools in their long-term efforts to increase student achievement on standardized tests" (p. 80). To address this issue, this systematic review is designed to highlight the strengths and weaknesses of quantitative research on block scheduling. Unfortunately, there is a major gap in this area, as there have been few "state of the art" experimental studies of block scheduling (Nichols, 2000; Stanley & Gifford, 1998).

Methodology

This systematic review followed a rigorous methodological approach informed by the WWC standards and prior research syntheses. The following section provides a step-by-step description of the search, retrieval, and coding processes. In addition, the statistical meta-analysis conducted on the evidence base is detailed.

Search Strategy

In the first stage of the systematic review of block scheduling research, a comprehensive search strategy was formulated. As the topic is primarily educational in nature, it was decided that the databases for Educational Resources Information Center (ERIC), Cambridge Scientific Abstracts, Dissertation Abstracts International (DAI), and OVID would be accessed to identify appropriate literature on this topic. To search these academic databases, the following key terms were used to construct Boolean logic statements: "block scheduling, alternating day block scheduling, full block scheduling, hybrid block scheduling, flexible scheduling, time block scheduling, extended class periods, 90-minute classes, school scheduling, 4X4 block scheduling, student achievement and block scheduling, and academic

achievement and block scheduling." Furthermore, the search was bounded to studies completed from 1993–2003, so that all data in the evidence report were collected and analyzed during the past ten years.

Search Results

The intensive electronic search yielded 292 documents consisting of journal articles, conference papers, evaluation reports, dissertations, and policy papers on block scheduling. The citations for the 292 documents were downloaded into Reference Manager 9, which is an interactive literature management software package.

During the second stage of the search process, abstracts for each of the 292 references were read and analyzed according to the initial criteria established for the systematic review. In reading the abstracts, 40 (14 percent) were reviewed by two members of the research team to ensure consistency in the acquisition decision. As a result, full-text copies were obtained for the 83 documents related to block scheduling interventions and outcomes within a high school setting. To ensure that no studies had been missed during the electronic search, a visual examination of the reference lists for each of the 83 articles/reports was conducted. After this manual search, 11 additional studies that met the inclusion criteria were obtained.

In addition, the Google and Yahoo search engines were accessed to locate additional online and published references that were not identified in the initial electronic or manual searches. The search of these Internet portals also was restricted to studies completed from 1993–2003. After this search, two additional studies that met the inclusion criteria were obtained, which qualified a total of 96 articles/reports for the third stage of review.

Keywording Criteria

During the third stage, a "keywording" rubric was used to categorize each study by the type of treatment, intensity of intervention, type of research design, type of sample, and type of outcome. To make it to the next phase, a study had to investigate a block scheduling treatment that was in place for more than one year using an experimental or quasi-experimental research design. Block scheduling is a complex, whole school intervention requiring that all teachers in a school change fundamental aspects of their teaching routines equally well. Hence, the evidence base ideally would include only studies where block scheduling had been in place for several years to allow teachers to maximize its effect, undisturbed by inefficiency in the day-to-day routines of implementation.

In addition, the sample for an eligible study had to be drawn from the high school setting and the outcomes of the research had to include specific student achievement measures (i.e., standardized tests). Furthermore, if there were multiple articles/reports from a single study, the most recent and/or complete document was eligible for inclusion. After being pilot tested for clarity and interrater reliability, the instrument was revised before the 96 articles/reports were coded and entered into a Microsoft Excel spreadsheet. Based on the aforementioned criteria, it was determined that 14 articles/reports were qualified to be included in the fourth stage of the review process. The most common reasons for exclusion were the type of research design (i.e., descriptive or qualitative study) and the type of outcome (e.g., GPA, student satisfaction).

Data Extraction

During the fourth stage of the review process, the 14 eligible studies were assessed on the quality of their research designs according to standards set forth by the What Works Clearinghouse Study Design and Implementation Device (DIAD). For the eight composite questions posed by the DIAD to be properly answered, a "data extraction" template that incorporated sub-questions related to validity issues and block scheduling characteristics was developed. Two researchers then coded each of the 14 studies and entered the results into an Excel spreadsheet. The researchers then came to consensus on the coding for each article/report. Finally, the 14 studies were assessed according to the eight composite questions in the DIAD. Nine of the articles/reports met the minimum inclusion criteria for all composite questions and were sent to the final stage of review. However, the What Works Clearinghouse released a new version (1.0) of the DIAD during this time, which necessitated an updating of the data extraction template and the re-extraction of the 14 studies.

Thus, a fifth stage of the review process was added, in which the researchers individually and as a team re-extracted each of the 14 articles/reports using the most recent version of the DIAD. After consensus was reached on the coding, the 14 studies were again evaluated relative to the eight composite questions. As a result, seven of the articles/reports fully met the minimum inclusion criteria and were sent to the sixth stage of review. Specifically, three studies that had previously qualified were subsequently deemed inadequate because of insufficient statistical reporting for effect size calculations. Additionally, one study that had not qualified was now considered to be appropriate for inclusion in the evidence report.

Meta-Analysis

According to Cohen (2001), "the concept behind meta-analysis is that different studies involving similar variables can be compared or combined by estimating the effect size in each study" (p. 237). Thus, a "standardized effect size" was computed using the Comprehensive Meta Analysis software program for the three most popular constructs in the evidence base, (i.e., mathematics, science, and English). As described by Borenstein and Rothstein (1999) in the preface to *Comprehensive Meta Analysis*, this software is a "stand alone program for meta-analysis used to synthesize data for multiple studies."

The most common statistics used in meta-analysis are Cohen's *d* and Hedges' *g*, both of which represent the strength of a relationship between the independent variable (e.g., traditional vs. block scheduling) and the dependent variable (e.g., test scores) in standard deviation units. In this study, Hedges' *g* was computed by dividing the difference between the group test means by the population pooled standard deviation estimate of the two groups. An effect size may be positive or negative, depending upon the direction of the difference. Because effect sizes are estimates and not parameters, confidence intervals are also provided in a meta-analysis to quantify some of the uncertainty inherent in capturing the "true" effect of an intervention. Most researchers suggest the following gauges for a *d* or *g* effect size: < .20 is small, .50 is medium, and > .80 is large (e.g., Cohen, 1988). From a practical perspective, a small effect size is interpreted as having little or no consequence for school practice or student outcomes, whereas medium and large effect sizes are interpreted as having more substantial implications for schools and students.

Overall effect sizes and confidence intervals were computed for mathematics, science, and English constructs. The overall effect size was computed as a weighted mean of the effect size for each measure with the weight for each study being the inverse of the square of the standard error. Thus, a study is given greater weight for a larger sample and more precise measurement, both of which reduce error.

Findings

After presenting the conclusions on the depth, breadth, and consistency of the block scheduling evidence base, the findings for each study are reported and compared in regard to the directionality and statistical significance of the block scheduling effect on student achievement. Several effect size analyses then are detailed with a focus on the magnitude and con-

sistency of block scheduling effects on scores from mathematics, English, and science tests. In addition, brief descriptions of the sample, treatment, measure, and results for each study are summarized in Appendix A.

CREAD Ratings

Before empirical findings were analyzed, the evidence base was assessed according to the eight composite questions posed in the Cumulative Research Evidence Assessment Device (CREAD). As stated in the CREAD, the objective is to "provide an expression of the confidence with which a conclusion can be drawn about the existence of causal effects of an intervention based on an entire body of accumulated evidence." Thus, the seven block scheduling studies were evaluated on their construct, internal, external, and statistical conclusion validity as defined by the CREAD. Because the CREAD conclusions provided sufficient confidence in the quality of the research designs, the results from the evidence base were appropriate for analysis.

Composite Question #1: The researchers are *confident* that the intervention was properly defined. As for construct validity, four of the seven studies fully reflected commonly held or theoretically derived ideas about the block scheduling intervention. Although three of the studies did not fully reflect shared ideas about block scheduling, their findings were consistent with the four well-defined studies. Specifically, the majority of results from both sets of studies indicated no statistically significant differences between treatment and comparison groups along with small negative effect sizes for the block scheduling intervention. In the studies that did not merit a "yes" on this composite question, either important details were missing from the description of the intervention and its implementation (e.g., instructional schedule) or the intervention was described only as a member of a broader class of block scheduling treatments.

Composite Question #2: According to the CREAD guidelines, the researchers are *somewhat unconfident* that the outcome measures were properly defined. Only one of the seven studies provided adequate evidence of construct validity for the outcome measures considered in this evidence report. In the studies that did not merit a "yes" on this question, there was evidence that the achievement tests had face validity and were properly aligned to the intervention. However, there was no evidence presented to suggest that the measures were reliable for measuring student achievement in block scheduled classes (e.g., internal consistency, interrater reliability), even though the standardized tests and statewide assessments previously have been shown to be psychometrically sound.

Composite Question #3: The researchers are *somewhat unconfident* that the participants in the group receiving the intervention were comparable to the participants in the comparison group. As for the internal validity of selection in the evidence base, none of the seven studies used groups that were comparable as defined by the DIAD (i.e., random assignment). Specifically, the seven studies were quasi-experimental, with quality ranging from weak to strong designs. However, each study employed adequate equating procedures to make the groups comparable (e.g., use of covariates) and there were no indications of severe overall or differential attrition.

Composite Question #4: The researchers are *confident* that the studies were free of events that happened concurrently with the intervention that confused its effect. Although the studies were all ex post facto, there was no positive evidence of contaminating events and no identified processes or events that were alternative explanations for the treatment effect.

Composite Question #5: The researchers are *somewhat confident* that the intervention was tested for its effectiveness using targeted participants, settings, outcomes, and occasions. As for the external validity of sampling, none of the studies merited a "yes" on this composite question, although most aspects of the theoretical population and common variations of settings, classes of outcomes, and data collection occasions were represented in the evidence base. For example, students with different ability levels were included in the evidence base as demonstrated by the variety of testing measures used in the studies (e.g., AP tests, statewide assessments). There were several studies, however, that focused on one school or one school district and thus only included a limited range of the important characteristics of the target population and settings.

Composite Question #6: The researchers are *somewhat confident* that the intervention was tested for its effectiveness within important subgroups of targeted participants, settings, outcomes, and occasions. As for the external validity of testing within subgroups, only one of the studies in the evidence base tested block scheduling for its effectiveness on all targeted subgroups. Furthermore, only a few studies in the evidence base tested for the time of measurement or for variations of block scheduling implementation. However, effect sizes can be estimated for a reasonable range of participants, settings, outcomes, and occasions in the other studies.

Composite Question #7: The researchers are *confident* that the studies allow for a precise estimation of effect size. As for the statistical conclusion validity for effect size estimation, all seven studies were based on statistical properties that allowed for valid estimates of effect sizes. Additionally, both sample sizes and the reliability of outcome measures were adequate to provide a sufficiently precise estimate of effect sizes.

Composite Question #8: The researchers are *confident* that studies were not systematically excluded because of their results. As for the statistical conclusion validity for completeness of reporting in the evidence base, the seven studies clearly did not censor data at the outcome level because a majority of the findings were of no statistically significant difference between groups. Furthermore, the literature search is presumed to have been effective in uncovering studies with limited availability, as dissertations and unpublished reports outnumber journal articles in the evidence base.

Overall Results

As displayed in Appendix A, four of the seven studies in the evidence base (i.e., Brake, 2000; Hackmann, Hecht, Harmston, Pliska, & Ziomek, 2001; Schreiber, Veal, Flinders, & Churchill, 2001; Walker, 2000) reported no statistically significant differences between traditional and block scheduling groups on all measured outcomes. The balance of the studies (i.e., McCreary & Hausman, 2001; Rice, Croninger, & Roellke, 2002; The College Board [TCB], 1998) reported mostly negative effects for the block scheduling intervention. To further explore trends across all seven studies, Appendix B summarizes the findings according to the main and subgroup effects of block scheduling on mathematics, science, English, and history test scores.

Mathematics Achievement

Six of the seven studies in the evidence base conducted analyses of student achievement on mathematics tests. As for the main effect results (i.e., block scheduling group vs. traditional scheduling group), three of the seven studies (Brake, 2000; Schreiber et al., 2001; Walker, 2000) found no statistically significant differences between schedule type and student achievement. The other three studies (McCreary & Hausman, 2001; Rice et al., 2002; TCB, 1998) reported statistically significant differences in favor of students in traditional scheduling (i.e., negative effects) on mathematics test scores. It should be noted that the TCB (1998) study found positive effects for students in A/B block scheduling but negative effects for students in 4X4 block scheduling after controlling for PSAT/NMSQT scores. The study conducted by Schreiber et al. (2001) provides the only subgroup analysis for a mathematics outcome measure. Specifically, they found no significant interactions between gender and schedule type or GPA group and schedule type.

English Achievement

As displayed in Appendix B, three of the seven studies conducted analyses of student achievement on English exams. Two of the studies (Brake, 2000; Schreiber et al., 2001) found no effect between scheduling type and student achievement while the TCB (1998) study yielded a negative effect for block scheduling groups in this content area. The College Board (1998), however, reported no effect for students in A/B block scheduling and a negative effect for students in 4X4 block scheduling based on adjusted test scores. Schreiber et al. (2001) conducted subgroup analyses and found no significant interactions between gender and schedule or GPA group and schedule type.

Science Achievement

Only two of the seven studies in the evidence base conducted analyses on science achievement tests. McCreary and Hausman (2001) found positive effects for students in block scheduling while the TCB (1998) study reported negative effects on science achievement. Again, The College Board (1998) found positive effects for students in A/B block scheduling but negative effects for students in 4X4 block scheduling after controlling for PSAT/NMSQT scores.

Other Subjects

Only one of the seven studies in the evidence base used history test scores as an outcome variable. Based on unadjusted test scores, the TCB (1998) study found a negative effect on history achievement for students in block scheduling. The College Board (1998) reported no effect for students in A/B block scheduling and a negative effect for students in 4X4 block scheduling after controlling for PSAT/NMSQT scores. The results for the Hackmann et al. (2001) study are reported in the "Other" column of Appendix B because the outcome measure was the ACT Composite Test, which is comprised of scores from mathematics, reading, English, and science reasoning subject areas. Hackmann et al. (2001) found no statistically significant differences between the treatment and comparison groups and found no effects in subgroup analyses for school size, urbanicity, and gender.

Primary Effect Size Analysis

The use of "vote counting" to discern the directionality of main and subgroup effects in the evidence base is a useful approach for empirically situating the block scheduling intervention. However, until effect size calculations are considered, there is no sense of the magnitude of the differences, regardless of statistical significance, between block scheduling

treatments and traditional scheduling approaches. For example, if the effect size values are similar across studies in the block scheduling evidence base, then the results of significance tests within studies can be dismissed (Cohen, 2001). However, effect size calculations are sensitive to differences in data reporting, sampling units, and outcome measures (see Appendix C for special considerations given to each study).

As displayed in Table 1, all of the effect sizes for all of the constructs, except for one positive science effect in the McCreary and Hausman (2001) study, were between –.052 and –.245. According to Cohen (1988), all of these effect sizes are considered to be "small" and negative for the block scheduling intervention. The seventh study in the evidence base (i.e., Rice et al., 2002) found a "medium" negative effect for block scheduling. Specifically, the regression coefficient for the class length variable was –.246, which translates to a Cohen's *d* of .50. However, this study did not provide standard errors, which precluded its inclusion into the primary and sensitivity effect size analyses.

Table 1
Descriptive Statistics and Effect Sizes for Block Scheduling Evidence Base

		Block			Traditional			
Study	Measure	*M*	*SD*	*n*	*M*	*SD*	*n*	*g*
Brake (2000)	ACT English	21.08	5.02	102	21.35	5.22	109	–.053
	ACT Math	20.03	3.21	102	20.71	5.26	109	–.155
The College Board (1998)	AP Biology	3.07	1.30	10756	3.27	1.28	34652	–.160
	AP English	2.94	1.09	18931	3.12	1.06	84467	–.168
	AP History	2.63	1.19	15194	2.91	1.20	83287	–.234
	AP Math	2.77	1.30	14263	2.88	1.29	60018	–.088
Hackmann et al. (2001)	ACT Composite	21.19	1.34	217	21.28	1.94	351	–.052
McCreary & Hausman (2001)	SAT Math	712.8	35.6	4800	717.8	35.6	2500	–.142
	SAT Science	693.5	36.9	4800	689.4	36.9	2500	.110
Schreiber et al. (2001)	ISTEP Math	68.14	20.11	175	70.59	17.45	142	–.129
	ISTEP English	65.29	16.67	175	69.44	17.27	142	–.245
Walker (2000)	Kansas Math	39.04	5.75	130	39.61	5.75	215	–.099

There was insufficient data on subgroup analyses regarding student characteristics, so effect sizes were not calculated in this area. For a more detailed analysis of each construct, Tables

2, 3, and 4 provide results for the effect sizes in mathematics, science, and English. Out of the six studies from the evidence base included in the effect size analysis, five considered mathematics achievement, three focused on English achievement, and two used science test scores as an outcome variable.

As displayed in Table 2, the combined effect size for the mathematics construct is –.095 with a lower and upper bound of –.112 and –.078, respectively. It is interesting to note that the three studies which reported no effect actually had small negative effects that were only non-significant because of small sample sizes. When these studies were combined with the two larger studies that also had small negative effects, the result is a statistically significant effect size that is negative in direction and small in magnitude.

Table 2
Effect Size Results for Mathematics Outcomes

Measure	Block (*n*)	Traditional (*n*)	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)
AP Math	14263	60018	–.088	–.106	–.070	.000
ACT Math	102	109	–.154	–.426	.118	.263
ISTEP Math	175	142	–.129	–.351	.094	.254
Kansas Math	130	215	–.099	–.318	.119	.371
SAT Math	4800	2500	–.142	–.190	–.094	.000
Combined	19470	62984	–.095	–.112	–.078	.000

As shown in Table 3, the overall effect size for the English construct is –.168 with a lower and upper bound of –.184 and –.152 respectively. As with mathematics, the combined and individual effect sizes for the English construct are negative and small.

Table 3
Effect Size Results for English Outcomes

Measure	Block (*n*)	Traditional (*n*)	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)
AP English	18931	84467	–.168	–.184	–.152	.000
ACT English	102	109	–.053	–.324	.219	.703
ISTEP English	175	142	–.245	–.467	–.021	.031
Combined	19208	84718	–.168	–.184	–.152	.000

As displayed in Table 4, the overall effect size for the science construct is –.115 with a lower and upper bound of –.134 and –.095, respectively. Although the science construct is home to the only positive effect size in the evidence base, the combined effect size for the two studies in this area is consistent with the "small negative" findings from the mathematics and English achievement outcomes.

Table 4
Effect Size Results for Science Outcomes

Measure	Block (*n*)	Traditional (*n*)	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)
AP Biology	10756	34652	–.160	–.181	–.138	.000
SAT Science	4800	2500	.110	.061	.158	.000
Combined	15556	37152	–.115	–.134	–.095	.000

When calculating the overall effect size of a construct, a *Q* statistic is computed to test for homogeneity of effect sizes for each study in the group. Specifically, the *Q* statistic indicates whether the variability in effect sizes is due to sampling error or some unmeasured variable(s), in which case the overall effect size is not reliably estimating the common population effect size. For this analysis, *Q* values were generated for the mathematics and English constructs. This was not done for science achievement because there were only two studies analyzing this construct.

As displayed in Table 5, the *Q* value for mathematics is 4.447, which is not statistically significant. Thus, the distribution of effect sizes in the five studies is homogeneous. Even though the tests used for mathematics were different, the effect sizes for the five studies are very close. This same pattern holds for the *Q* value in English, as it is a non-significant 1.167. A *t* value is also computed to test the significance of the overall effect size by dividing the absolute value of the mean effect size by the standard error of the mean effect size (Lipsey & Wilson, 2001). Specifically, this is a two-tailed test of the null hypothesis that the overall effect size is not significantly different from zero. For mathematics, $t = –11.03$; $p < .001$, which indicates that the small negative effect size is statistically significant. For English, $t = –20.99$; $p < .001$, which indicates that the small negative effect size is statistically significant.

Table 5
Combined Effect Size Statistics

	Test of Null Hypothesis				Test of Heterogeneity		
Outcome	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)	*Q* Value	*df*	*p* value
Mathematics	–.095	–.112	–.078	.000	4.447	4	.349
English	–.168	–.184	–.152	.000	1.167	2	.558

Sensitivity Effect Size Analysis

Although different block scheduling plans were combined for the primary effect size analysis, a sensitivity analysis also was conducted to compare 4X4 and A/B interventions with traditional scheduling. In addition to the Rice et al. (2002) study, this analysis resulted in the loss of the Walker (2000) study because data were not disaggregated by schedule type. Furthermore, the trimester group from the McCreary and Hausman (2001) study and the hybrid group from the Schreiber et al. (2001) study were not included in the sensitivity analysis. Thus, the McCreary and Hausman (2001) study was included only in the A/B analysis and the Schreiber et al. (2001) study was included only in the 4X4 analysis. A final sensitivity analysis was conducted to compare studies in which there were both 4X4 and A/B groups. Specifically, the effect sizes from the TCB (1998), Brake (2000), and Hackmann et al. (2001) studies were compared in this fashion.

Tables 6, 7, 8, and 9 provide the results of the sensitivity analysis that compared 4X4 block scheduling with traditional scheduling. As displayed in Table 6, all of the effect sizes for all of the constructs, except for one positive effect on the ACT composite test in the Hackmann et al. (2001) study, were between –.020 and –.359. All of these effect sizes are considered to be small and negative for 4X4 block scheduling.

Table 6

Descriptive Statistics and Effect Sizes in 4X4 Block Scheduling Interventions

		4X4			Traditional			
Study	Measure	*M*	*SD*	*n*	*M*	*SD*	*n*	*g*
Brake (2000)	ACT English	21.50	4.80	59	21.60	5.00	63	–.020
	ACT Math	20.20	3.90	59	21.30	5.20	63	–.237
The College Board (1998)	AP Biology	2.89	1.36	2853	3.27	1.28	34652	–.299
	AP English	2.99	1.10	9732	3.12	1.06	84467	–.122
	AP History	2.48	1.17	5490	2.91	1.20	83287	–.359
	AP Math	2.64	1.30	5279	2.88	1.29	60018	–.186
Hackmann et al. (2001)	ACT Composite	21.36	1.09	56	21.28	1.94	351	.043
Schreiber et al. (2001)	ISTEP Math	70.08	21.28	49	70.59	17.45	142	–.027
	ISTEP English	66.92	17.44	49	69.44	17.27	142	–.145

As shown in Table 7, the combined effect size for the mathematics construct is –.185 with a lower and upper bound of –.213 and –.157, respectively. Similar to the primary analysis, the result is a statistically significant effect size that is negative in direction and small in magnitude for the 4X4 block scheduling intervention as compared with traditional scheduling.

Table 7

Effect Size Results for Mathematics Outcomes in 4X4 Block Scheduling Interventions

Measure	Block (*n*)	Traditional (*n*)	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)
AP Math	5279	60018	–.186	–.214	–.158	< .05
ACT Math	59	63	–.154	–.593	.120	> .05
ISTEP Math	49	142	–.027	–.352	.297	> .05
Combined	5387	60223	–.185	–.213	–.157	< .05

As displayed in Table 8, the overall effect size for the English construct is –.122 with a lower and upper bound of –.143 and –.101, respectively. Similar to the primary analysis, the result is a statistically significant effect size that is negative in direction and small in magnitude for the 4X4 block scheduling intervention as compared with traditional scheduling.

Table 8

Effect Size Results for English Outcomes in 4X4 Block Scheduling Interventions

Measure	Block (*n*)	Traditional (*n*)	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)
AP English	9732	84467	–.122	–.143	–.101	< .05
ACT English	59	63	–.020	–.375	.335	> .05
ISTEP English	49	142	–.145	–.470	.181	> .05
Combined	9840	84672	–.122	–.143	–.101	< .05

As displayed in Table 9, the *Q* value for mathematics is .006, which is not statistically significant. Thus, the distribution of effect sizes in the three 4X4 studies is homogeneous. This same pattern holds for the *Q* value in English, as it is a non-significant .336. As for the null hypothesis tests, the results for both mathematics and English indicate that the small negative effect sizes for the 4X4 block scheduling intervention are statistically significant.

Table 9

Combined Effect Size Statistics for 4X4 Block Scheduling Interventions

	Test of Null Hypothesis				Test of Heterogeneity		
Outcome	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)	*Q* Value	*df*	*p* value
Mathematics	–.185	–.213	–.157	< .05	.006	2	> .05
English	–.122	–.143	–.101	< .05	.336	2	> .05

Tables 10, 11, 12, 13, and 14 provide the results of the sensitivity analysis that compared A/B block scheduling with traditional scheduling. As shown in Table 10, all of the effect sizes for all of the constructs, except for one positive effect on the SAT science test in the McCreary and Hausman (2001) study, were between –.024 and –.220. All of these effect sizes are considered to be small and negative for A/B block scheduling.

Table 10

Descriptive Statistics and Effect Sizes for A/B Block Scheduling Interventions

		A/B			Traditional			
Study	Measure	*M*	*SD*	*n*	*M*	*SD*	*n*	*g*
Brake (2000)	ACT English	20.50	5.30	43	21.00	5.50	46	–.092
	ACT Math	19.80	2.30	43	19.90	5.40	46	–.024
The College Board (1998)	AP Biology	3.13	1.27	7903	3.27	1.28	34652	–.110
	AP English	2.89	1.08	9199	3.12	1.06	84467	–.220
	AP History	2.72	1.20	9704	2.91	1.20	83287	–.158
	AP Math	2.84	1.31	8984	2.88	1.29	60018	–.031
Hackmann et al. (2001)	ACT Composite	21.13	1.41	161	21.28	1.94	351	–.084
McCreary & Hausman (2001)	SAT Math	714.3	35.6	2400	717.8	35.6	2500	–.098
	SAT Science	694.2	36.9	2400	689.4	36.9	2500	.130

As shown in Table 11, the combined effect size for the mathematics construct is –.040 with a lower and upper bound of –.060 and –.019, respectively. Similar to the main analysis, the result is a statistically significant effect size that is negative in direction and small in magnitude for the A/B block scheduling intervention as compared with traditional scheduling.

Table 11

Effect Size Results for Mathematics Outcomes in A/B Block Scheduling Interventions

Measure	Block (*n*)	Traditional (*n*)	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)
AP Math	8984	60018	–.031	–.053	–.009	< .05
ACT Math	43	46	–.024	–.439	.392	> .05
SAT Math	2400	2500	–.098	–.154	–.042	< .05
Combined	11427	62564	–.040	–.060	–.019	< .05

As displayed in Table 12, the overall effect size for the English construct is –.220 with a lower and upper bound of –.241 and –.198, respectively. Similar to the primary analysis, the result is a statistically significant effect size that is negative in direction and small in magnitude for the A/B block scheduling intervention.

Table 12

Effect Size Results for English Outcomes in A/B Block Scheduling Interventions

Measure	Block (*n*)	Traditional (*n*)	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)
AP English	9199	84467	–.220	–.238	–.195	< .05
ACT English	43	46	–.092	–.508	.324	> .05
Combined	9242	84513	–.220	–.241	–.198	< .05

As shown in Table 13, the overall effect size for the science construct is –.071 with a lower and upper bound of –.093 and –.049, respectively. The combined effect size for the A/B block scheduling intervention is consistent with the small negative finding from the primary effect size analysis.

Table 13

Effect Size Results for Science Outcomes in A/B Block Scheduling Interventions

Measure	Block (*n*)	Traditional (*n*)	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)
AP Biology	7903	34652	–.110	–.134	–.085	< .05
SAT Science	2400	2500	.130	.074	.186	> .05
Combined	10303	37152	–.071	–.093	–.049	< .05

Combined effect size statistics were not computed for the English and science constructs because there were only two studies in each. As displayed in Table 14, the *Q* value for mathematics is 4.750, which is not statistically significant. Thus, the distribution of effect sizes in the three A/B studies is homogeneous. As for the null hypothesis test, the results for mathematics indicate that the small negative effect size is statistically significant.

Table 14

Combined Effect Size Statistics for A/B Block Scheduling Interventions

	Test of Null Hypothesis				Test of Heterogeneity		
Outcome	Effect (*g*)	Lower (*CI*)	Upper (*CI*)	Sig. (*p*)	*Q* Value	*df*	*p* value
Mathematics	–.040	–.060	–.019	< .05	4.750	2	> .05

It is clear from the sensitivity analysis that the block scheduling intervention, be it 4X4 or A/B, results in a small negative effect on student achievement in mathematics and English as compared to traditional scheduling. Although the magnitude of the negative effect is larger for the 4X4 group on the mathematics construct, the A/B group has a considerably larger negative effect for the English construct.

The final sensitivity analysis compared the 4X4 and A/B groups from the three studies in which data were disaggregated for both types of block scheduling plans. No conclusive findings emerged, as the 4X4 group had a significant, albeit small, advantage over the A/B group on the ACT English, Mathematics, and Composite tests while the A/B group held the same small advantage on the AP Biology, History, and Mathematics tests. The one interesting finding is that the 4X4 group had a slight advantage on the AP English test, in addition to the advantage on the ACT English test and the overall smaller negative effect on the English construct.

Table 15
Descriptive Statistics and Effect Sizes for Studies with 4X4 and A/B Groups

		4X4			A/B			
Study	Measure	*M*	*SD*	*n*	*M*	*SD*	*n*	*g*
Brake (2000)	ACT English	21.50	4.80	59	20.50	5.30	43	.198
	ACT Math	20.20	3.90	59	19.80	2.30	43	.120
College Board (1998)	AP Biology	2.89	1.36	2853	3.13	1.27	7903	–.189
	AP English	2.99	1.10	9732	2.89	1.08	9199	.092
	AP History	2.48	1.17	5490	2.72	1.20	9704	–.202
	AP Math	2.64	1.30	5279	2.84	1.31	8984	–.153
Hackmann et al. (2001)	ACT Composite	21.36	1.09	56	21.13	1.41	161	.172

Discussion

Although the systematic review and synthesis of the block scheduling literature has answered some important questions about this educational intervention, many new questions have been raised in the process. The following conclusions, limitations, recommendations, and implications are designed to accurately and concisely represent the findings from this

evidence report while providing a context and framework for future conceptual and applied work in this area.

Conclusions

The findings for this research synthesis paint two slightly different pictures of the effect of block scheduling on high school student achievement. The first picture is less than crystal clear, in that there was a predominance of non-significant findings for the seven studies along with inconsistency regarding the positive and negative effects reported in the evidence base. For example, findings for main effects across subject areas were consistently neutral for English and decidedly mixed for mathematics and science.

However, the second picture is more understandable, as there was a consistent effect size pattern within and across studies in three construct domains analyzed. Specifically, studies with statistically significant findings indicated a small negative effect, studies with non-significant results showed a small negative effect, and the overall findings of the evidence base demonstrated a small negative effect of block scheduling on student achievement in mathematics, English, and science. In this synthesis, meta-analytic statistical techniques provided a more definitive answer by uncovering the underlying direction and magnitude of block scheduling effects for studies that reported no effect because of small sample sizes (Cohen, 2001). Thus, what first seemed to be inconclusive turned out to be more consistent and clear regarding the effect of block scheduling on high school student achievement. Furthermore, the results from the sensitivity analysis bolster this conclusion, in that a small negative effect on student achievement was present for both 4X4 and A/B interventions as compared with traditional scheduling.

Although this conclusion is in conflict with Rettig and Canady's (2001) contention that, "block scheduling will not have a negative effect on student achievement" (p. 81), there are likely no harmful effects of block scheduling on high school student achievement as measured by test scores. The reason is that the magnitude of the effect sizes is small enough to have little or no practical consequences. The limited results from the subgroup analyses also support this assertion, in that no student groups were reported to be adversely affected by block scheduling. In addition, these results should be considered in light of the consistency in non-academic findings for block scheduling studies, the theoretical support for the block scheduling intervention, and the relative cost-effectiveness of block scheduling implementation.

Limitations

A variety of methodological and implementation limitations have prevented researchers from conducting appropriate quantitative analyses regarding the effect of block scheduling on student achievement outcomes (Rettig & Canady, 2001). Furthermore, Wronkovich (1998) expresses skepticism about the objectivity of research on block scheduling when he chides adherents and opponents alike for being "nearly evangelical in their zeal to promote their position" (p. 3).

As evidenced by the small number of studies that qualified for inclusion in the evidence base, the major limitation encountered in this systematic review was the relatively weak standing of research on this topic. It was planned that two or three studies considered to be "best evidence" would be examined as a further check on the consistency of findings in the evidence base. Unfortunately, the lack of depth in the block scheduling literature resulted in no studies being identified as such when the DIAD standards were used as the criteria. Specifically, a lack of confidence regarding the comparability of groups for the ex post facto quasi-experimental studies combined with other design flaws limited the tenability of this systematic review. In addition, several studies that would have qualified for a best evidence designation (e.g., Zhang, 2000) did not make it to the evidence base because of missing effect size information.

There were also challenges confronted during the effect size analysis that limit the usefulness of the calculations. First, some studies reported on a small sample of students from a few schools (e.g., Brake, 2000), while others employed a much larger data set of students (e.g., TCB, 1998). As a result, studies with large sample sizes essentially eliminated "outliers" in the data. Thus, a small difference between means with a large sample may give a statistically significant result while not translating to a large effect size. Second, the unit of measurement for sample size varies somewhat across the studies in the evidence base. Although most researchers used *N* to represent students, a few used *N* to represent schools (i.e., Hackmann et al., 2001; Walker, 2000). The consequence of this limitation is that studies using aggregated data at the school level are under-weighted in the effect size analysis when pooled with studies that have sample sizes in the thousands of students. Third, the student samples in the evidence base were drawn from different populations, in that students taking the ACT and AP tests are more likely to be higher achieving than are students taking statewide tests like the ISTEP. Unfortunately, the sparse evidence base precluded disaggregating these studies to conduct a differential analysis based on student ability. However, the results were consistent across all studies regardless of the student population analyzed.

One limitation that may lead to an underestimation of block scheduling effects on student achievement is the psychometric weaknesses of most standardized achievement tests. To the extent that the outcome measures of the studies in the evidence base were oriented to measuring facts instead of critical thinking, there was a misalignment between the nature of the intervention and the fullest representation of its effects. This misalignment also may account for the disparity in findings between the associated qualitative and quantitative research, in that surveys and observations are more sensitive to the benefits of block scheduling than are test scores. Nonetheless, the strong focus on high-stakes testing at the federal and state levels necessitated the use of these outcome measures for this systematic review.

A final limitation is that block scheduling is more of a vehicle to enact instructional, curricular, and organizational reform than a full-fledged educational intervention. Specifically, "the adoption of a block schedule typically is part of a larger school improvement effort with many other programs or policies being simultaneously implemented" (Rettig & Canady, 2001, p. 82). Thus, the effect of block scheduling on student achievement might be more difficult to detect. Furthermore, there has been a recent wave of hybrid block schedules that are even more challenging to analyze (Rettig & Canady, 2001). This limitation provides support for researching block scheduling only after it has been implemented for several years so that the treatment can be consistently integrated into the fabric of a school or classroom.

Recommendations

Based on the conclusions generated from this systematic review, the following recommendations are grouped into suggestions for block scheduling stakeholders and educational researchers. If the goal of a block scheduling intervention is to raise the short-term test scores of students, the evidence base does not support such an approach. However, if the goal is to impact non-academic outcomes while improving test scores in the long term, the findings from the evidence base and research literature are more supportive. For block scheduling to remain a viable option in the educational reform repertoire, researchers must work more closely with stakeholders to design programs that have greater research validity.

As for future research, there is a need for better designed studies, appropriate statistical analyses, and sensitive and diverse outcome measures to accurately gauge the effect of block scheduling on student achievement. Specifically, experimental research approaches are necessary if more studies are to be included in the block scheduling evidence base. Furthermore, few quantitative studies of block scheduling have been sufficiently longitudinal to provide reliable data from which to make comparisons and draw conclusions (Stanley & Gifford,

1998). Although an ANCOVA is a legitimate statistical method for controlling group differences in a quasi-experimental study, there are other techniques (e.g., Hierarchical Linear Modeling) that offer promise in this area. Another area for study is the development of achievement tests in all content areas that measure the learning affected by block scheduling (e.g., cooperative, experiential).

There also is a need to break out the effects of block scheduling across significant and important subgroups of target participants, settings, outcomes, occasions, and intervention variations. In a prior study in a junior high school setting, a large and positive effect size on language arts and science test scores was found for lower-achieving students in block scheduling programs (Lewis et al., 2003). Furthermore, the interaction between schedule type and teacher quality could be investigated in regard to its relationship with student achievement. Teachers implementing block scheduling could be randomly assigned from within a school to eliminate differential teacher quality effects of those using block scheduling versus those using traditional scheduling. Finally, qualitative studies that explore why block scheduling works or does not work are a natural outgrowth of this evidence report.

Implications

The most compelling implications of this systematic review are for the WWC Evidence Report rubric. First and foremost, dissertations, unpublished reports, and conference papers were rated more highly than journal articles because of the ability to include information on effect sizes, instrument reliability, sampling design, and other methodological criteria. This seemingly unintended consequence has the potential to change how educational researchers and journal publishers interact and collaborate.

Although the DIAD is well constructed and relatively easy to navigate, there are several inconsistencies between the study DIAD and the coding guide that may cause uncertainty during the data extraction process. Furthermore, for such a systematic and prescribed process, the composite questions in the CREAD seem to rely too much on arbitrary benchmarks. For example, the small number of studies in the block scheduling evidence base caused confidence statements for the design quality to be quite sensitive to the addition or subtraction of just one study. As evidenced by the results of this research synthesis, all future systematic reviews should be required to generate effect size calculations. This would certainly strengthen the conclusions from these reviews and may inspire researchers to more regularly report the necessary data to conduct meta-analyses.

At this time, block scheduling does not merit consideration as a subject for a full WWC Evidence Report. However, if future research better captures the full range of interventions,

outcomes, and samples for block scheduling, this topic may become more appropriate for a larger systematic review.

Bibliography

Borenstein, M., & Rothstein, H. (1999). *Comprehensive meta-analysis*. Englewood, NJ: Biostat Inc.

Brake, N. L. (2000, November). *Student course-taking delivered through a high school block schedule: The relationship between the academic core and student achievement*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Buckman, D. C., King, B. B., & Ryan, S. (1995). Block scheduling: A means to improve school climate. *NASSP Bulletin, 79*(571), 9–18.

Cawelti, G. (1994). *High school restructuring: A national survey*. Arlington, VA: Educational Research Service.

Cohen, B. H. (2001). *Explaining psychological statistics* (2nd ed.). New York: John Wiley & Sons, Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Deuel, L-L. S. (1999). Block scheduling in large, urban high schools: Effects on academic achievement, student behavior, and staff perceptions. *The High School Journal, 83*(1), 14–25.

Edwards, C. M., Jr. (1995). Virginia's 4X4 high schools: High school, college, and more. *NASSP Bulletin, 79*(571), 23–41.

Eineder, D. V., & Bishop, H. L. (1997). Block scheduling the high school: The effects on achievement, behavior, and student-teacher relationships. *NASSP Bulletin, 81*(589), 45–54.

Hackmann, D. G., Hecht, J. E., Harmston, M. T., Pliska, A. M., & Ziomek, R. L. (2001, April). *Secondary school scheduling models: How do types of models compare to the ACT scores?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Hamdy, M., & Urich, T. (1998). Perceptions of teachers in south Florida toward block scheduling. *NASSP Bulletin, 82*(596), 79–82.

Knight, S. L., DeLeon, N. J., & Smith, R. G. (1999). Using multiple data sources to evaluate an alternative scheduling model. *The High School Journal, 83*(1), 1–13.

Lapkin, S., Harley, B., & Hart, D. (1997). Block scheduling for language study in middle grades: A summary of the Carleton Case Study. *Learning Languages, 2*(3), 4–8.

Lewis, C. W., Cobb, R.B., Winokur, M., Leech, N., Viney, M. & White, W. (2003, November 11). The effects of full and alternative day block scheduling on language arts and science achievement in a junior high school. *Education Policy Analysis Archives, 11*(41). Retrieved October 1, 2003, from http://epaa.asu.edu/epaa/v11n41/.

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

McCreary, J., & Hausman, C. (2001). *Differences in student outcomes between block, semester, and trimester schedules*. (ERIC Document Reproduction Service No. ED457590)

Nichols, J. D. (2000). Scheduling reform: A longitudinal exploration of high school block scheduling structures. *International Journal of Educational Reform, 9*, 134–147.

O'Neil, J. (1995). Finding time to learn. *Educational Leadership, 53*(3), 11–15.

Pisapia, J., & Westfall, A. L. (1997). *Alternative high school scheduling: A view from the teacher's desk. Research report* (Report No. UD031866). Richmond, VA: Metropolitan Educational Research Consortium. (ERIC Document Reproduction Service No. ED 411335)

Queen, J. A., Algozzine, B., & Eaddy, M. (1997). Implementing 4X4 block scheduling: Pitfalls, promises, and provisos. *NASSP Bulletin, 81*(588), 107–114.

Rettig, M. D., & Canady, R. L. (2001). Block scheduling: More benefits than challenges. Responses to Thomas (2001). *NASSP Bulletin, 85*(628), 78–86.

Rice, J. K., Croninger, R. G., & Roellke, C. F. (2002). The effect of block scheduling high school mathematics courses on student achievement and teachers' use of time: Implications for educational productivity. *Economics of Education Review*, 21, 599–607.

Schreiber, J. B., Veal, W. R., Flinders, D. J., & Churchill, S. (2001, November 14). Second year analysis of a hybrid schedule high school. *Education Policy Analysis Archives*, 9(46). Retrieved June 26, 2003, from http://epaa.asu.edu/epaa/v9n46/.

Shortt, T. L., & Thayer, Y. (1995). What can we expect to see in the next generation of block scheduling? *NASSP Bulletin, 79*(571), 53–62.

Skrobarcek, S. A., Chang, H-W. M., Thompson, C., Johnson, J., Atteberry, R., Westbrook, R., & Manus, A. (1997). Collaboration for instructional improvement: Analyzing the academic impact of a block scheduling plan. *NASSP Bulletin, 81*(589), 104–111.

Stanley, A., & Gifford, L. J. (1998). *The feasibility of 4X4 block scheduling in secondary schools: A review of the literature*. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Staunton, J. (1997). A study of teacher beliefs on the efficacy of block scheduling. *NASSP Bulletin, 81*(593), 73–80.

Stewart, J. W., & Shank, J. (1971). Daily demand of modular flexible scheduling for small schools. *Educational Leadership, 29*, 537–544.

Thayer, Y. V., & Shortt, T. L. (1998). Block scheduling can enhance school climate. *Educational Leadership, 56*(4), 76–81.

The College Board. (1998, May). *Block schedules and student performance on AP examinations* (Research Notes No. RN-03). New York: Author.

Thomas, C., & O'Connell, R. W. (1997). *Parent perceptions of block scheduling in a New York State public high school* (Report No. EA028507). New York: Educational Research Organization. (ERIC Document Reproduction Service No. ED 409 644).

Walker, G. (2000). The effect of block scheduling on mathematics achievement in high and low SES secondary schools (Doctoral dissertation, University of Kansas, 2000). *Dissertation Abstracts International,* 61, 4638.

Wallinger, L. M. (2000). The effect of block scheduling on foreign language learning. *Foreign Language Annals, 33*(1), 36–50.

Weller, D. R., & McLeskey, J. (2000). Block scheduling and inclusion in a high school. *Remedial and Special Education, 21*, 209–218.

Wilson, J. W., & Stokes, L. C. (1999). Teachers' perceptions of the advantages and measurable outcomes of the 4X4 block scheduling design. *The High School Journal, 83*(1), 44–54.

Wilson, J. W., & Stokes, L. C. (2000). Students' perceptions of the effectiveness of block versus traditional scheduling. *American Secondary Education, 28*(3), 3–12.

Wronkovich, M. (1998). Block scheduling: Real reform or another flawed educational fad? *American Secondary Education, 26*(4), 1–6.

Zhang, G. (2000, April). *Academic performance differences between students in block and traditionally scheduled high schools 1993–2000*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Appendix A: Matrix of Studies Included in the Evidence Base

Author and Publication Date	Sample	Treatment	Measures	Results
Brake (2000)	288 students from the graduating class of 1995 and 2000 at two high schools	4X4 semester block schedule and A/B block schedule	ACT tests in Mathematics and English	No statistically significant difference between traditional and block scheduling groups on either test
Hackmann, Hecht, Harmston, Pliska, & Ziomek (2001)	568 high schools (with 38,089 seniors) from the states of Illinois and Iowa in 1999	4X4 semester block schedule and A/B block schedule	ACT Composite Test	No statistically significant difference between traditional and block scheduling groups on test
McCreary & Hausman (2001)	28,526 students in three high schools during the 1995–1999 school years	Trimester schedule and A/B block schedule	Stanford Achievement Test (SAT9) in Mathematics and Science	Block students had significantly lower total mathematics scores and significantly higher total science scores
Rice, Croninger, & Roellke (2002)	12,000 tenth-grade students from 1,200 high schools during 1990 NELS follow-up	Mathematics courses with over 70 minutes of class time	NELS:88 Achievement Test Battery in Mathematics	Enrollment in a block-scheduled course had a statistically significant but negative impact on mathematics scores
Schreiber, Veal, Flinders, & Churchill (2001)	318 sophomores from one high school in 1998	4X4 semester block schedule and hybrid block schedule	ISTEP tests in Reading, Language, and Mathematics	No statistically significant differences between traditional and block scheduling groups on tests
The College Board (1998)	Students completing the four highest volume AP examinations in 1997 (range from 26,238 to 64,300)	4X4 semester block schedule and A/B block schedule	Advanced Placement (AP) tests in English Literature, U.S. History, Biology and Calculus	4X4 and A/B block scheduling students had significantly lower math, science, history, and English scores than traditional students before controlling for PSAT/NMSQT scores
Walker (2000)	345 high schools (with 29,514 students) from Kansas during the 1994–1999 school years	4X4 semester block schedule and A/B block schedule	Kansas State Mathematics Assessment	No statistically significant difference between traditional and block scheduling groups on test

Appendix B: Summary of Findings for Main and Subgroup Effects by Subject Area

Author and Publication Date	Mathematics	Science	English	Other
Brake (2000)	*Main:* No effect	N/A	*Main:* No effect	N/A
Hackmann et al. (2001)	N/A	N/A	N/A	*Main:* No effect *Subgroup:* No effects for school size, urbanicity, gender, years on block schedule
McCreary & Hausman (2001)	*Main:* Negative effect	*Main:* Positive effect	N/A	N/A
Rice et al. (2002)	*Main:* Negative effect	N/A	N/A	N/A
Schreiber et al. (2001)	*Main:* No effect *Subgroup:* No effect for gender or GPA group	N/A	*Main:* No effect *Subgroup:* No effect for gender or GPA group	N/A
The College Board (1998)	*Main:* Negative effect for 4X4 and A/B students*	*Main:* Negative effect for 4X4 and A/B students*	*Main:* Negative effect for 4X4 and A/B students*	*Main:* Negative effect for 4X4 and A/B students*
Walker (2000)	*Main:* no effect	N/A	N/A	N/A

No effect - No statistically significant difference ($p > .05$) between treatment and comparison groups

Positive effect - Statistically significant difference ($p < .05$) in favor of block scheduling group

Negative effect - Statistically significant difference ($p < .05$) in favor of traditional scheduling group

*College Board reported different findings based on an analysis with PSAT/NMSQT scores as a covariate

Appendix C: Special Considerations for Effect Size Calculations

In the Brake (2000) study, the sample size for the traditional group represents students at two high schools on a traditional schedule that completed the ACT English and ACT Mathematics tests in 1995. The sample size for the block group represents students at the same two high schools that completed the ACT English and ACT Mathematics tests in 2000. This sample is considered the block group because each high school implemented a block scheduling format during the 1996–2000 school years. As the high schools implemented different types of block scheduling, students from the 2000 cohort were combined for the primary effect size analysis but separated into 4X4 and A/B groups for the sensitivity analysis.

In the Hackmann et al. (2001) study, the sample size for the traditional group represents the number of high schools on traditional schedules in the sample drawn from Illinois and Indiana, whereas the sample size for the block group represents the number of high schools using 4X4 and A/B plans. However, high schools from the sample were separated for the sensitivity analysis based on the type of block scheduling plan. The ACT Composite score was used as an outcome measure, which precluded the incorporation of this study in the effect size analysis for the mathematics, English, and science constructs. However, the ACT Composite still makes sense as a dependent variable because it represents a commonly used high-stakes testing outcome.

For the McCreary and Hausman (2001) study, students from one high school using a traditional schedule comprised the traditional group while students from two high schools formed the block scheduling group. Again, the block scheduling group contains students exposed to two types of the treatment, as data were combined for A/B schools and for schools on a trimester schedule (i.e., five classes a day at 63 minutes per period). For the sensitivity analysis, only schools on the A/B block scheduling plan were included. To calculate effect sizes, eta squared values (i.e., the percent of variance between the groups accounted for by scheduling type) were converted to *g* scores by using the *F* values from an ANCOVA analysis (see Cohen, 1988). However, unadjusted means were not reported in this study, which necessitated using the adjusted means for the effect size calculation. Furthermore, only the pooled standard deviation was provided, so an assumption of equality of variances was made in order to complete the effect size analysis using the available data.

For the Schreiber et al. (2001) study, the sample size for the block scheduling group represents students on both a 4X4 schedule and a hybrid schedule in the same high school. However, only students participating in the 4X4 plan were included in the sensitivity analysis. The effect size analysis used data from the Indiana Statewide Testing for Educational Progress (ISTEP) "Mathematics Total" scores, while the analysis for English used combined data from the ISTEP "Reading Total" and "Language Total" scores.

In The College Board (1998) study, the sample size for the traditional group includes students on traditional schedules that completed the AP exam for Biology, English, History, and/or Mathematics in 1997. The sample size for the block scheduling group contains students from two types of block schedules (i.e., 4X4 and A/B) who completed the same AP exams in 1997. The sensitivity analysis compared students from each type of block schedule with students on traditional schedules from the same time period. Effect sizes were computed using data from the full sample because the standard errors reported for a smaller sample that included only students with PSAT/NMSQT scores to serve as a covariate resulted in standard deviations that were not reliable.

In the Walker (2000) study, the unit of allocation was schools, so the sample sizes represent the number of high schools that used a traditional or a block schedule each year during a five-year period. The block scheduling group is comprised of schools that used 4X4 and A/B plans. This study was not included in the sensitivity analysis because the data were not disaggregated by the type of block scheduling plan.

As for the Rice et al. (2002) study, the data were not included in the either the primary or sensitivity effect size analyses, because means and standard deviations could not be computed from the information provided in the study. However, the study is included in the evidence base because the coefficients from the Hierarchical Linear Modeling analysis are interpretable as effect sizes.