APPENDIX

THE CONSISTENT OUTCOME OF BILINGUAL EDUCATION PROGRAMS

A Meta-Analysis of Meta-Analyses

Grace P. McField
California State University San Marcos

David R. McField
University of Southern California

Paper presented at the American Educational Research Association, April 2007, Chicago, IL and at the International Symposium on Bilingualism, May 2007, Hamburg, Germany.

Author Note: This project was made possible through a grant from the American Educational Research Association and Institute for Education Sciences, U.S. Department of Education.

The author would like to acknowledge and thank Marie Thomas, Professor of Psychology, California State University, San Marcos, for her invaluable insights and assistance with technical and meta-analytic issues in the development of this paper.

The Miseducation of English Learners, pages 267–297
Copyright © 2014 by Information Age Publishing
All rights of reproduction in any form reserved.

ABSTRACT

This meta-analysis provides a snapshot of the major bilingual education meta-analyses, and reports the findings of an innovative approach to considering both program and research quality in quantitative bilingual education reviews. First, a review of meta-analyses in the literature is provided, showing that bilingual education meta-analyses conducted independently and examining different studies have consistently reached similar conclusions. Second, primary studies drawn from the pool of previous reviews are reanalyzed, with attention to both program quality (strong, light, weak and undefined bilingual education programs) and research quality, and effect sizes calculated. The findings reveal that considering both program quality and research quality in evaluating outcomes of bilingual education programs renders a very different outcome than considering research quality alone. Specifically, when both program quality and research quality were considered, there was a higher effect size than when only research quality was considered, with nearly double the magnitude found for the former. In this study, the inclusion of program quality factors resulted in an effect size of d = .41 vs. an effect size of d = .26 when only research quality was factored into calculations.

INTRODUCTION AND BACKGROUND

The American public is under the impression that bilingual education doesn't work. Yet even a quick glance at the professional literature shows that it does. Study after study has reported that children in bilingual programs typically outperform their counterparts in all-English programs on tests of academic achievement in English. Or, at worst, they do just as well on tests in English. Moreover, bilingual education programs provide other benefits such as biliteracy and bicultural/multicultural development.

Numerous reviews of the research literature have confirmed the conclusion that bilingual education works. Recent reviews include those conducted by Rolstad, Mahoney, and Glass (2005) and Slavin and Cheung (2005), as well as Francis, Lesaux and August (2006), a report originally sponsored by but not released by the U.S. Department of Education.[1] All three found an advantage for bilingual education. For scientists—and, one would hope, for policymakers—it is highly significant when reviews of the literature, conducted independently and examining different studies, reach similar conclusions. Such consistency provides strong evidence that research findings are reliable, rather than merely the result of chance.

It is also noteworthy that the latest reviews used a sophisticated methodology that is considered more precise and more objective than earlier approaches to summarizing research findings. The methodology is known as *meta-analysis*.

Until recently, most reviews o
described as "narrative" or "vote-
ies, decide which ones are worthy
as favoring either bilingual or al
each study—regardless of how bi
comes, how many subjects are inv
ods—gets one vote. Then the vo
winner declared.

Several reviews of this kind ha
more effective than all-English p
English and to progress academ
1978; Cummins, 1983; Krashen, 1
Kanter (1981) concluded there w
bilingual education. Alone amor
(1996) counted more studies favo
also reported only small difference
the existence of high-quality biling
has been systematically refuted in
odological rigor and findings (e.g.,

Meta-analysis, by contrast, allow
sive approach. Using powerful st
numerous variables in each study,
student and teacher characteristic
duration of study, year of publicatio
peer-reviewed journal), and so fort
subjectivity, sometimes called "revi
or in deciding which studies to excl

Perhaps most important, meta-a
to measure *effect size*—how big an
demonstrates over another—expre
or overall effect size can be then c
taking into account the degree of pc
for each primary study.

Other advantages of effect sizes i
index that can be compared across
index that cuts across different test
used to inform practice and policy. T
reach general conclusions about the
cal approach versus another. It has b
represents a small impact of a treat
impact and .80 represents a large i
interpreted to roughly equal two, fi
bilingual education programs (Cun

ACT

of the major bilingual education me-
an innovative approach to consider-
y in quantitative bilingual education
in the literature is provided, showing
conducted independently and exam-
reached similar conclusions. Second,
previous reviews are reanalyzed, with
ng, light, weak and undefined bilin-
n quality, and effect sizes calculated.
h program quality and research qual-
education programs renders a very
earch quality alone. Specifically, when
ity were considered, there was a high-
quality was considered, with nearly
ormer. In this study, the inclusion of
fect size of d = .41 vs. an effect size of
actored into calculations.

D BACKGROUND

npression that bilingual education
at the professional literature shows
orted that children in bilingual pro-
terparts in all-English programs on
ish. Or, at worst, they do just as well
l education programs provide other
al/multicultural development.
terature have confirmed the conclu-
ecent reviews include those conduct-
005) and Slavin and Cheung (2005),
2006), a report originally sponsored
ment of Education.[1] All three found
or scientists—and, one would hope,
ant when reviews of the literature,
ing different studies, reach similar
s strong evidence that research find-
e result of chance.
reviews used a sophisticated meth-
ise and more objective than earlier
indings. The methodology is known

Until recently, most reviews of bilingual education research have been described as "narrative" or "vote-counting." Scholars collect a body of studies, decide which ones are worthy of inclusion, and characterize each study as favoring either bilingual or all-English programs. In narrative reviews, each study—regardless of how big a difference it finds in educational outcomes, how many subjects are involved, or how rigorous its research methods—gets one vote. Then the votes are counted for each approach and a winner declared.

Several reviews of this kind have concluded that bilingual education is more effective than all-English programs in helping children to acquire English and to progress academically (Zappert and Cruz, 1977; Troike, 1978; Cummins, 1983; Krashen, 1996). On the other hand, Baker and de Kanter (1981) concluded there was no advantage (but also no harm) to bilingual education. Alone among narrative reviews, Rossell and Baker (1996) counted more studies favoring all-English programs, although they also reported only small differences between treatments and acknowledged the existence of high-quality bilingual programs. Rossell and Baker (1996) has been systematically refuted in the literature for both issues with methodological rigor and findings (e.g., Greene, 1998 and 1999).

Meta-analysis, by contrast, allows reviewers to take a more comprehensive approach. Using powerful statistical techniques, it can control for numerous variables in each study, including sample size, program model, student and teacher characteristics, research design, outcome measures, duration of study, year of publication, type of publication (e.g., dissertation, peer-reviewed journal), and so forth. These techniques can also minimize subjectivity, sometimes called "reviewer bias," in characterizing outcomes or in deciding which studies to exclude or include.

Perhaps most important, meta-analysis gives reviewers the opportunity to measure *effect size*—how big an advantage one educational treatment demonstrates over another—expressed as a single number. A grand total or overall effect size can be then calculated for the studies under review, taking into account the degree of positive or negative effect sizes calculated for each primary study.

Other advantages of effect sizes include the fact that it is a standardized index that can be compared across studies. The effect size is a consistent index that cuts across different tests and background factors that can be used to inform practice and policy. Thus meta-analysis makes it possible to reach general conclusions about the relative effectiveness of one pedagogical approach versus another. It has been suggested that an effect size of .20 represents a small impact of a treatment, while .50 represents a modest impact and .80 represents a large impact (Cohen, 1977). This has been interpreted to roughly equal two, five, and eight months' advantage for bilingual education programs (Cummins, 2000). According to another source, the National Institute of Education's Joint Dissemination Review Panel (Tallmadge, 1977), for the field of education, .33 sd = educationally significant, and in some cases, .25 sd = educationally significant. There are also fail-safe calculations that can be done to see how many studies with negative outcomes would need to be located in order to render the average positive effect size null.

Reviewing the Reviews

This section is a "*meta*-meta-analysis," a summary of the findings of published meta-analyses of programs for English language learners (ELLs). The intent herein is to determine how much confidence should be placed in these reviews and what overall conclusions we should draw from them.

Eight major reviews (seven meta-analyses and one narrative review by Demmert and Towner, 2003) have compared the two broad program types of bilingual and all-English programs. Despite slightly different criteria for including studies and different dates of publication, the average effect sizes across the majority of these reviews are remarkably similar, with students in bilingual education showing consistently positive outcomes when compared to those in all-English classrooms as follows.[2]

Review	N	Dates	Mean ES
Willig (1985)	23	1971–1980	0.33
Greene (1997)	11	1972–1991	0.18
McField (2002)	10	1968–1985	0.28
Rolstad et al. (2005)	17	1985–	0.23
Slavin & Cheung (2005)	17	1971–	0.33
Demmert and Towner (2003)	2	1982–1988	1.12
Okada et al. (1982)	168	1965–1980	0.13–0.24
Oh (1987)	54	1984–1987	1.21

Note: N = number of studies included in meta-analysis
ES = effect size

Some caveats are in order. With the exception of Demmert and Towner (2003), all of these reviews examined studies conducted in the United States only and lasting for about one academic year or about nine to ten months. Demmert and Towner (2003) included primary studies that examined bilingual education programs in Australia (Murtagh, 1982), and arctic Canada (Wright, Taylor, and Macarthur, 2000), although the latter could not be included in the set of studies for which effect sizes were calculated due to study limitations. However, one year may not be enough time for bilingual programs to show their positive effects. Additionally, in most

studies reviewed in the meta-ar
tal (bilingual) students were E
comparison students were flu
stringent comparison in repor

That said, the findings of tl
all consistently positive, rangi
that the findings of the five m
primary studies included in the
Greene, 1997; McField, 2002; I
have been consistently positive,
.18 to .33. (Note: Okada et al. 1
the mean effect size calculation
sizes were calculated for studies
one non-meta-analytic review in
with a mean of 1.12.

In all studies included in the
ucation programs were compa
Two of the meta-analyses (Willig
of vote-counting reviews (Bake
1996). Three others (McField,
Slavin and Cheung, 2005) used
studies for review.

In addition to the foregoing m
is also included in this review. Al
not a meta-analysis per se, this
gual Native American language p
of bilingual programs for this c
population. Another meta-analys
but no primary studies from this
tion due to the fact that no break
included. Similarly, no primary s
the present meta-analysis due to
the primary studies used measur
ing and other factors related to
studies from Oh (1987) were incl

There are, of course, wide var
ing from dual language to early-e
options. There are also wide vari
only, some allowing a small amou
simply "submersing" children in t
lengths to make sure English inpu
variations were considered and in
made comparisons between biling

ion's Joint Dissemination Review
education, .33 sd = educationally
lucationally significant. There are
ne to see how many studies with
ted in order to render the average

ı summary of the findings of pub-
ıglish language learners (ELLs).
uch confidence should be placed
ions we should draw from them.
yses and one narrative review by
ıred the two broad program types
spite slightly different criteria for
ublication, the average effect sizes
remarkably similar, with students
ıtly positive outcomes when com-
s follows.[2]

Dates	Mean ES
1971–1980	0.33
1972–1991	0.18
1968–1985	0.28
1985–	0.23
1971–	0.33
1982–1988	1.12
1965–1980	0.13–0.24
1984–1987	1.21

neta-analysis

xception of Demmert and Town-
studies conducted in the United
ademic year or about nine to ten
included primary studies that ex-
ı Australia (Murtagh, 1982), and
ırthur, 2000), although the latter
s for which effect sizes were calcu-
one year may not be enough time
sitive effects. Additionally, in most

studies reviewed in the meta-analyses, comparison students and experimental (bilingual) students were ELLs. But in some studies that were included, comparison students were fluent speakers of English, making for a more stringent comparison in reports of student outcomes.

That said, the findings of the seven meta-analyses and one review were all consistently positive, ranging from .18 to 1.21. Noteworthy is the fact that the findings of the five meta-analyses that included effect sizes for all primary studies included in the review (listed chronologically, Willig, 1985; Greene, 1997; McField, 2002; Rolstad et al. 2005; Slavin & Cheung, 2005) have been consistently positive, with a mean effect size of .26 and a range of .18 to .33. (Note: Okada et al. 1982 and Oh, 1987 could not be included in the mean effect size calculation due to study limitations. See below.) Effect sizes were calculated for studies found in Demmert and Towner (2003), the one non-meta-analytic review included herein, and ranged from .84—1.17, with a mean of 1.12.

In all studies included in these meta-analyses, students in bilingual education programs were compared with students in all-English programs. Two of the meta-analyses (Willig, 1985, and Greene, 1999) were re-analyses of vote-counting reviews (Baker and de Kanter, 1981; Rossell and Baker, 1996). Three others (McField, 2002; Rolstad, Mahoney, and Glass, 2005; Slavin and Cheung, 2005) used their own criteria in selecting a group of studies for review.

In addition to the foregoing meta-analyses, Demmert and Towner (2003) is also included in this review. Although Demmert and Towner (2003) was not a meta-analysis per se, this valuable review examined studies of bilingual Native American language programs and helped to expand the review of bilingual programs for this culturally and linguistically diverse student population. Another meta-analysis, Okada et. al (1982) was also reviewed, but no primary studies from this review could be included in the next section due to the fact that no breakdown of the individual primary studies was included. Similarly, no primary studies from Oh (1987) were included in the present meta-analysis due to the fact that the bulk of the tests used in the primary studies used measures that could not be confirmed for norming and other factors related to reliability and validity. Thus, no primary studies from Oh (1987) were included in the present meta-analysis.

There are, of course, wide variations among bilingual programs, ranging from dual language to early-exit, to late-exit to concurrent translation options. There are also wide variations among programs labeled English-only, some allowing a small amount of help in the primary language, some simply "submersing" children in the mainstream, and some going to great lengths to make sure English input is comprehensible for ELLs. Many such variations were considered and included in this review, so long as studies made comparisons between bilingual and all-English programs.

It could be argued, of course, that the similar mean effect sizes across the different meta-analyses is due to the fact that the meta-analyses featured many of the same studies and were simply redundant. To determine whether this was the case, studies reviewed in more than one meta-analysis were examined (Table 1). Most comparisons were tests of reading comprehension in English, although a small number of studies of other measures of English proficiency was also used (e.g., oral measures were used in Skoczylas, 1972; and in Murtagh, 1982). Further, comparisons in which fluent English speakers served as comparison students were excluded. This method not only allowed us to determine overlap, but also served as a way of measuring reliability, that is, to see whether different researchers came up with similar results.

Table 1 shows that, while there is some overlap, it is clear that all investigators did not examine the same body of primary research studies. The vast majority of studies appeared in only one or two of the five meta-analyses. So there was broad support for results favoring bilingual education.

On the other hand, when studies did appear in more than one review, there was substantial agreement about their effect size, even though effect sizes can be calculated in different ways that can produce different results. The only serious disagreement involved the effect size calculated for Saldate et al. (1985), but in all three meta-analyses the effect size was positive.

What Kind of Bilingual Program?

In the meta-meta-analysis above, a deliberate attempt was made to look at the big picture to see whether there was general agreement among studies. Individual meta-analyses have focused on different aspects in conducting reviews of bilingual education.

Willig (1985) analyzed a number of methodological variables, reporting that studies using random assignment of subjects to experimental and comparison groups resulted in higher effect sizes favoring bilingual education. Greene (1997) reported a similar pattern. Willig also found that when comparison groups contained elements of bilingual education, such as significant use of the native language, the advantage for the bilingual program was weaker. When comparison groups contained students who had exited the bilingual program, the effect size in favor of bilingual education was considerably lower (d = –.03, versus d = .38). Willig concluded that positive effects for bilingual education were apparent only when methodological weaknesses in the studies were controlled. In other words, the tighter the research design, the stronger the effects for bilingual education.

Others have investigated the impact of the kind of bilingual program used. McField (2002) concluded that programs designed along principles

TABLE 1 Comparison of Studies of Reading Comprehension Included in Previous Meta-Analyses[3]

	Slavin & Cheung (2005)	Willig (1985)	Greene (1997)	McField (2002)	Demmert & Towner (2003)	Rossell & Kuder (2005)	Rolstad et al. (2005)
Alvarez (1975)	–0.23					–.05	
Huzar (1978)	0.31		0.18	.31, .01		0.16	
Plante (1976)	0.5		0.52			0.52	
Ramirez et al. (1991)			0.12			0.26	0.01

the similar mean effect sizes across
the fact that the meta-analyses fea-
ere simply redundant. To determine
wed in more than one meta-analysis
parisons were tests of reading com-
all number of studies of other mea-
used (e.g., oral measures were used
982). Further, comparisons in which
parison students were excluded. This
ine overlap, but also served as a way
whether different researchers came

me overlap, it is clear that all investi-
of primary research studies. The vast
e or two of the five meta-analyses. So
oring bilingual education.
id appear in more than one review,
their effect size, even though effect
s that can produce different results.
ed the effect size calculated for Sal-
-analyses the effect size was positive.

?

leliberate attempt was made to look
was general agreement among stud-
sed on different aspects in conduct-

of methodological variables, report-
ent of subjects to experimental and
effect sizes favoring bilingual educa-
pattern. Willig also found that when
s of bilingual education, such as sig-
advantage for the bilingual program
contained students who had exited
in favor of bilingual education was
= .38). Willig concluded that positive
pparent only when methodological
lled. In other words, the tighter the
ts for bilingual education.
ct of the kind of bilingual program
programs designed along principles

TABLE 1 Comparison of Studies of Reading Comprehension Included in Previous Meta-Analyses[3]

	Slavin & Cheung (2005)	Willig (1985)	Greene (1997)	McField (2002)	Demmert & Towner (2003)	Rossell & Kuder (2005)	Rolstad et al. (2005)
Alvarez (1975)	–0.23					–.05	
Huzar (1973)	0.31		0.18	.31, .01		0.16	
Plante (1976)	0.5		0.52			0.52	
Ramirez et al. (1991)			0.12			0.25	0.01
Campeau et al. (1975) Corpus Christi	0.45					0.45	
Maldonado (1994)[4]	1.66					0.12	
Campeau et al. (1975) Alice	0.49					0.45	
Saldate et al. (1985)	0.89			0.42		1.47	1.47
Morgan (1971)	0.26			0.26		0.27	
Carter & Chatfield (1986)							.32
Doebler & Mardis (1980)	0.15					0.15	
Covey (1973)	0.72	0.74	0.74	0.74		0.66	
Medrano (1986, 1988)							.10, –.18
Kaufman (1968)	0.23	0.31	0.2	.49, .11		0.2	
Rothfarb, Ariza, Urrutia (1987)							
Danoff et al. (1977)		0.01	–0.12			0.12	

(continued)

TABLE 1 Comparison of Studies of Reading Comprehension Included in Previous Meta-Analyses[3] (continued)

	Slavin & Cheung (2005)	Willig (1985)	Greene (1997)	McField (2002)	Demmert & Towner (2003)	Rossell & Kuder (2005)	Rolstad et al. (2005)
McSpadden (1979)		0.2					
Olesini (1971)		0.97					
Stebbins et. al. (1977)		–0.06					
Stern (1975)		–0.48					
Lindholm (1991)							–0.59
Medina, Saldate & Mishra (1985)				–0.22, –0.13, –0.51			–0.3, –.57
Texas Education Agency (1988)							–0.06
Powers (1978)			–0.33	–0.44		– .35	
Rossell (1990)			–0.05			–.25	
Bacon et al. (1982)			0.68	0.82, 0.98	N/R*	0.7	
Cohen (1975)	0					–.21, .08, –.28	
Cottrell (1971)[5]					N/R		
Franks (1988)[6]					N/R		
Murtagh (1982)					N/R		

* *N/R:* Study included but no effect size reported

hypothesized to underlie ideal
were more effective. But very f
one "strong" program and fou
this way). Rolstad, Mahoney, anc
that late-exit or developmental b
early-exit or transitional prograr
the research base on studies tha
education (CBE) on academic o
Their review included studies t
along with primary language ins
the small base of available quali
for CBE, they concluded that "t
erature on CBE programs for Na

The present review and analys
ysis that focused on program qua
meta-analyses in the field that c
into one big pool, the differentia
of varying program quality (stro
examined. In addition a grand n
of bilingual education across all
comparison. This way, the averag
grams that were of acceptable re
with the average effect size for bi
acceptable research quality by p
and undefined).

HYI

The following hypotheses were fo
sidering both program quality and
of bilingual education. (Note: No
both program quality and researc

1. For studies of both acceptabl
the better the bilingual educ
bilingual education program
(reported in effect sizes).
2. For studies of both acceptabl
students in undefined bilingu
strate weak effect sizes relativ
bilingual education programs

			[illegible]			
Texas Education Agency (1988)						−0.06
Powers (1978)		−0.33	−0.44		− .35	
Rossell (1990)		−0.05			−.25	
Bacon et al. (1982)		0.68	0.82, 0.98	N/R*	0.7	
Cohen (1975)	0				−.21, .08, −.28	
Cottrell (1971)[5]				N/R		
Franks (1988)[6]				N/R		
Murtagh (1982)				N/R		

* *N/R*: Study included but no effect size reported

hypothesized to underlie ideal bilingual programs (e.g., Krashen, 1996) were more effective. But very few such comparisons were possible (only one "strong" program and four "weak" programs could be analyzed in this way). Rolstad, Mahoney, and Glass (2005) present evidence suggesting that late-exit or developmental bilingual programs are more effective than early-exit or transitional programs. Demmert and Towner (2003) reviewed the research base on studies that examined the effects of culturally based education (CBE) on academic outcomes among Native American students. Their review included studies that combined culturally based education along with primary language instruction (i.e., bilingual education). While the small base of available qualitative studies were found to show support for CBE, they concluded that "the availability of quantitative research literature on CBE programs for Native American children is severely limited."

The present review and analysis expands on McField's (2002) meta-analysis that focused on program quality. As in McField (2002), unlike previous meta-analyses in the field that categorized all bilingual education studies into one big pool, the differential impact of bilingual education programs of varying program quality (strong, light, weak and undefined) was also examined. In addition a grand mean effect size or average overall impact of bilingual education across all program quality levels was computed for comparison. This way, the average effect size for bilingual education programs that were of acceptable research quality only, could be compared with the average effect size for bilingual education programs that were of acceptable research quality by program quality level (strong, light, weak and undefined).

HYPOTHESES

The following hypotheses were formulated to test the interaction of considering both program quality and research quality in quantitative reviews of bilingual education. (Note: No meta-analysis in the field has considered both program quality and research quality other than McField, 2002.)

1. For studies of both acceptable and unacceptable research quality, the better the bilingual education program (strong, light, or weak bilingual education programs), the better the students' outcomes (reported in effect sizes).
2. For studies of both acceptable and unacceptable research quality, students in undefined bilingual education programs will demonstrate weak effect sizes relative to students in strong, light, and weak bilingual education programs.

3. The better the research quality (research design, control for bias, etc.) *and* program quality, the higher the effect size.

METHODOLOGY

Studies were selected from the previous major qualitative and quantitative reviews of bilingual education. In order to address the *file-drawer bias* issue (Wolf, 1986), unpublished studies (e.g., dissertations) were also included (see Table 2). Studies were reviewed and categorized for program quality as strong, light, weak, and undefined bilingual education programs as follows. According to Krashen (1996), there are three components of a strong bilingual education program: 1. Comprehensible input in English, typically in the form of ESL instruction (CI-ESL) at beginning levels; and comprehensible input in English in subject matter areas, typically sheltered instruction (CI-SM), at intermediate levels; 2. Literacy development or reading instruction in the L1 (L1-LIT); and 3. Subject matter teaching in the L1 (L1-SM). A study was categorized as a strong bilingual education program if it had all three components; light if it had two components, 1 & 2 or 1 & 3; and weak if it had one component, 2 or 3. A study was considered undefined if there was not enough information to determine the program quality.

Concerning research quality, studies were categorized as sound or acceptable if they met the following criteria. Similar criteria have been used in previous meta-analyses conducted by Francis, Lesaux, & August (2006), Greene (1998), Slavin & Cheung (2005), and Rossell and Baker (1996).

Five Characteristics of Acceptable Studies (Rossell & Baker, 1996, pp. 13–14)

1. They were true experiments in which students were randomly assigned to treatment and control groups;
2. They had non-random assignment that either matched students in the treatment and comparison groups on factors that influence achievement, or statistically controlled for them;
3. They included a comparison group of LEP students of the same ethnicity and similar language background;
4. Outcome measures were in English using normal curve equivalents (NCEs), raw scores, scale scores, or percentiles, but not grade equivalents;
5. There were no additional educational treatments, or the studies controlled for additional treatments if they existed.

Two additional criteria were used from Greene's (1998) meta-analysis:

6. Studies needed to have ad
mental group receiving so
and the control group rec
7. Sufficient control (randor
ences) was utilized for init
different IQs between the

Studies were categorized as f
the above criteria. After careful
egorized as follows:

- 11/23 strong bilingual pr
gram cohorts; and 5/23 w
undefined bilingual progr
- Concerning research qual
sound, while 10/15 studie

Next, effect sizes were calcula
ity category & for each research
ferent statistics were used to calc
ranging from Glass' d, Cohen's d
adjusted g, among others (Hed
and Wolf, 1986). For the presen
Hedges' original g. Hedges' orig
effect size for several reasons. Th
and a pooled variance, given tha
ondly, the strengths and weakn
most. Finally, Hedges' original
follow transformations of Hedge
For example, Rosenthal (1991) r
Hedge's original g to Cohen's d
Effect size measures were also
Kim and Grissom, 2005; Wolf, 19
a slightly biased estimator of effec
thal and Rubin (1982) provided
curate. Wolf (1986) reports that
a weighted average d = Σwd/Σw
estimator works well as long as th
effect size is not greater than 1.5.
tor of Rosenthal and Rubin (198
Hedges' g into Cohen's d using t
(1991) in an attempt to approxi
compile a summary effect size fo

arch design, control for bias,
r the effect size.

OGY

ıajor qualitative and quantitative
o address the *file-drawer bias* issue
dissertations) were also included
ategorized for program quality as
al education programs as follows.
ree components of a strong bilin-
ible input in English, typically in
eginning levels; and comprehen-
as, typically sheltered instruction
development or reading instruc-
atter teaching in the L1 (L1-SM).
al education program if it had all
ponents, 1 & 2 or 1 & 3; and weak
as considered undefined if there
the program quality.
ere categorized as sound or ac-
. Similar criteria have been used
rancis, Lesaux, & August (2006),
and Rossell and Baker (1996).

h students were randomly as-
ups;
hat either matched students
oups on factors that influence
ed for them;
of LEP students of the same
ground;
using normal curve equiva-
s, or percentiles, but not grade

al treatments, or the studies
if they existed.

eene's (1998) meta-analysis:

6. Studies needed to have adequate control groups, with the experimental group receiving some primary language (L1) instruction, and the control group receiving "English-only" instruction.
7. Sufficient control (random assignment, statistical control for differences) was utilized for initial differences such as initial test scores or different IQs between the bilingual program and control group.

Studies were categorized as flawed or unacceptable if they did not meet the above criteria. After careful review, the set of primary studies were categorized as follows:

- 11/23 strong bilingual program cohorts; 4/23 light bilingual program cohorts; and 5/23 weak bilingual program cohorts; and 3/23 undefined bilingual program cohorts.
- Concerning research quality, 5/15 studies were methodologically sound, while 10/15 studies were methodologically flawed.

-Next, effect sizes were calculated and compared for each program quality category & for each research quality category (see Table 3). Several different statistics were used to calculate effect sizes in previous meta-analyses, ranging from Glass' d, Cohen's d, Glass' g, Hedges' original g, and Hedges' adjusted g, among others (Hedges & Olkin, 1985; Kim & Grissom, 2005; and Wolf, 1986). For the present review, all effect sizes were calculated for Hedges' original g. Hedges' original g was used as the default estimator of effect size for several reasons. The first reason is that it uses sample means and a pooled variance, given that we used sample not population data. Secondly, the strengths and weaknesses of Hedge's original g are known to most. Finally, Hedges' original g is very transparent, in that it is easy to follow transformations of Hedges' original g from one metric to another. For example, Rosenthal (1991) re-presents his 1986 formula for converting Hedge's original g to Cohen's d transformation: $g = (N/df)^{1/2}$.

Effect size measures were also transformed into Cohen's d (Cohen, 1977; Kim and Grissom, 2005; Wolf, 1986). Hedges (1982) demonstrated that d is a slightly biased estimator of effect size, but both Hedges (1982) and Rosenthal and Rubin (1982) provided a method to make the effect size more accurate. Wolf (1986) reports that Rosenthal and Rubin's (1982) formula for a weighted average $d = \Sigma wd/\Sigma w$ where $w = 2N/8 + d^2$, and states that this estimator works well as long as the sample sizes are greater than 10 and the effect size is not greater than 1.5. In the present review the unbiased estimator of Rosenthal and Rubin (1982) was used after transforming the study Hedges' g into Cohen's d using the transformation provided by Rosenthal (1991) in an attempt to approximate an unbiased estimator and to also compile a summary effect size for each category (e.g., all strong bilingual

TABLE 2 Studies Included in the Present Meta-Analysis, with Comparison to Previous Meta-Analyses[7]

	Same as present meta-analysis McField (2007)[8]	Slavin & Cheung (2005)	Willig (1985)	Greene (1997)	McField (2002)	Demmert & Towner (2003)	Rossell & Kuder (2005)	Rolstad et al. (2005)
Alvarez (1975)		−0.23					–.05	
Huzar (1973)	d = .31, .01	0.31		0.18	.31, .01		0.16	
Maldonado (1994)[9]	d = 1.82	1.66					0.12	
Plante (1976)		0.5		0.52			0.52	
Ramirez et al. (1991)				0.12			0.25	0.01
Campeau et al. (1975) Corpus Christi	N/A statistical limitations	0.45					0.45	
Campeau et al. (1975) Alice	N/A statistical limitations	0.49					0.45	
Saldate et al. (1985)	study d = .15	0.89			0.42		1.47	1.47
Morgan (1971)	study d = .30	0.26			0.26		0.27	
Carter & Chatfield (1986)								.32
Doebler & Mardis (1980)	d = 0.15	0.15					0.15	
Covey (1973)	study d = .61	0.72	0.74	0.74	0.74		0.66	
Medrano (1986, 1988)								.10, –.18
Kaufman (1968)	d = .27, .20	0.23	0.31	0.2	0.49, 0.11		0.2	

(continued)

TABLE 2 Studies Included in the Present Meta-Analysis, with Comparison to Previous Meta-Analyses[7] (continued)

	Same as present meta-analysis McField (2007)[8]	Slavin & Cheung (2005)	Willig (1985)	Greene (1997)	McField (2002)	Demmert & Towner (2003)	Rossell & Kuder (2005)	Rolstad et al. (2005)
Rothfarb, Ariza, Urrutia (1987)	d = –.09 for Cohort I 2nd grade d = −0.46 Cohort II 1st grade							

Campeau et al. (1975) Alice	N/A statistical limitations	0.49					0.45	
Saldate et al. (1985)	study d = .15	0.89			0.42		1.47	1.47
Morgan (1971)	study d = .30	0.26			0.26		0.27	
Carter & Chatfield (1986)								.32
Doebler & Mardis (1980)	d = 0.15	0.15					0.15	
Covey (1973)	study d = .61	0.72	0.74	0.74	0.74		0.66	
Medrano (1986, 1988)								.10, –.18
Kaufman (1968)	d = .27, .20	0.23	0.31	0.2	0.49, 0.11		0.2	

(continued)

TABLE 2 Studies Included in the Present Meta-Analysis, with Comparison to Previous Meta-Analyses[7] (continued)

	Same as present meta-analysis McField (2007)[8]	Slavin & Cheung (2005)	Willig (1985)	Greene (1997)	McField (2002)	Demmert & Towner (2003)	Rossell & Kuder (2005)	Rolstad et al. (2005)
Rothfarb, Ariza, Urrutia (1987)	d = –.09 for Cohort I 2nd grade d = –0.46 Cohort II 1st grade							
Danoff et al. (1977)			0.01	–0.12			0.12	
McSpadden (1979)			0.2					
Olesini (1971)			0.97					
Stebbins et. al. (1977)			–0.06					
Stern (1975)			–0.48					
Lindholm (1991)								–0.59
Medina, Saldate & Mishra (1985)	study d = –.29				–0.22, –0.13, –0.51			–.3, –.57
Texas Education Agency (1988)	N/A statistical limitations							–0.06
Powers (1978)[10]	(not included in average ES calculations–outlier)			–0.33	–0.44[11]		– .35	

(continued)

TABLE 2 Studies Included in the Present Meta-Analysis, with Comparison to Previous Meta-Analyses[7] (continued)

	Same as present meta-analysis McField (2007)[8]	Slavin & Cheung (2005)	Willig (1985)	Greene (1997)	McField (2002)	Demmert & Towner (2003)	Rossell & Kuder (2005)	Rolstad et al. (2005)
Rossell (1990)				–0.05			–.25	
Bacon et al. (1982)	d = .82, .98			0.68	0.82, 0.98	N/R*	0.7	
Cohen (1975)		0					–.21, .08, –.28	
Cottrell[12] (1971)	d = .62					N/R		
Franks (1988)[13]	d = 1.34, .99, 1.28 for 3 cohorts					N/R		
Murtagh (1982)	d = 1.2, .47, 1.04					N/R		
Skoczylas (1972)	study d = .31							
De la Garza (1985)	study d = .17							

* *N/R:* Study included but no effect size reported

TABLE 3 Strong, Light, and Weak Bilingual Education Studies

Study/Exam	Quality of BE Program	Grade Level Tested	Length of Program	Time of Study	ES g (McField, 2002)	ES d (McField, 2002)	ES r (McField, 2002)	Sound/Flawed Methodology?	Greene (1998)	Rossell & Baker (1996)
1. Doebler & Mardis (1980) MAT—Reading	strong	2nd	7 months[14]	immediate	.15	.15		sound	not reviewed	?
2. Maldonado (1994) CTBS—Language & Reading	strong	2–4* 3–5*	2 years 2 years	graduates graduates				sound	not reviewed	?
		*combined single cohort			1.78	1.82				
3. Skoczylas (1972) Listening Comprehension	strong		2 years	immediate				sound		

	1.28 for 3 cohorts	
Murtagh (1982)	d = 1.2, .47, 1.04	N/R
Skoczylas (1972)	study d = .31	
De la Garza (1985)	study d = .17	

* *N/R:* Study included but no effect size reported

TABLE 3 Strong, Light, and Weak Bilingual Education Studies

Study/Exam	Quality of BE Program	Grade Level Tested	Length of Program	Time of Study	ES			Sound/Flawed Methodology?	Greene (1998)	Rossell & Baker (1996)
					g	d	r			
					(McField, 2002)					
1. Doebler & Mardis (1980)	strong	2nd	7 months[14]	immediate	.15	.15		sound	not reviewed	?
MAT—Reading										
2. Maldonado (1994)	strong	2–4*	2 years	graduates				sound	not reviewed	?
CTBS—Language &		3–5*	2 years	graduates						
Reading		*combined single cohort			1.73	1.82				
3. Skoczylas (1972)	strong		2 years	immediate				sound		
Listening Comprehension		1st	(K–1)		.16	.16	.08		.13 Rdg.[15]	TBE =
Oral		1st			.46	.47	.23		–.05Eng.[16]	submersion
						Study d = .31				
4. Cottrell (1971)	strong	1st	9 months	immediate	.61	.62		flawed	not reviewed	?
MAT—Total										
5. De La Garza et al. (1985)	strong									
Reading Comprehension	(SAT)	1st	1 year	immediate	.10	.10	.05	flawed	inadequate	not reviewed
	(CAT)	2nd	2 years		.19	.19	.10		control for	
	(CAT)	3rd	3 years		.21	.21	.10		differences	
						d = .17				
Reading Voc.	(SAT)	1st	1 year		–.32	–.32	–.16			
	(CAT)	2nd	2 years		.50	.50	.24			
	(CAT)	3rd	3 years		.25	.25	.12			
						d = .14				
						Study d = .16				

(continued)

TABLE 3 Strong, Light, and Weak Bilingual Education Studies (continued)

Study/Exam	Quality of BE Program	Grade Level Tested	Length of Program	Time of Study	ES			Sound/Flawed Methodology?	Greene (1998)	Rossell & Baker (1996)
					g	d	r			
					(McField, 2002)					
6. Franks (1988)	strong	1st	2 years[17]	immediate	1.32	1.34		flawed	not reviewed	?
		2nd	3 years		.99	.99				
		3rd	4 years		1.27	1.28				
					3 cohorts (avg d = 1.17)					
7. Medina et al. (1985) Total Reading	strong	1st–5th (5 years)	graduates							
		6th			–.21	**–.22**	–.11	flawed	not reviewed	not reviewed
		8th			–.13	–.13	–.07			
		12th			–.50	–.51	–.25			
					Study d = –.29					
8. Saldate et al. (1985)	strong	2nd	2 years	immediate	–.29	–.29	–.14	flawed	not reviewed	not reviewed
		3rd	3 years		**.91**	**.93**	**.42**			
					Study d = .15					
9. Rothfarb	strong	K			–.09	–.09		unacceptable	not reviewed	??
	light	1st			–.46	–.46				
					2 cohorts					
10. Murtagh (1982) Oral	light	1st	1 year[18]	immediate	1.13	1.20		flawed	not reviewed	?
		2nd	2 years	immediate	.45	.47				
		3rd	3 years	immediate	.98	1.04				
					3 cohorts (avg d = .89)					

(continued)

TABLE 3 Strong, Light, and Weak Bilingual Education Studies (continued)

Study/Exam	Quality of BE Program	Grade Level Tested	Length of Program	Time of Study	ES			Sound/Flawed Methodology?	Greene (1998)	Rossell & Baker (1996)
					g	d	r			
					(McField, 2002)					
11. Huzar (1973)	weak	2nd	2 years	immediate	.01	**.01**	.01	sound	.18 Rdg.	TBE = submersion
		3rd	3 years		.31	**.31**	.15		.18 Eng.	
					2 cohorts					
12. Morgan (1971)	weak		7 months	immediate						
1. Word Reading		1st			.37	.38	.19	flawed	not reviewed	TBE > submersion
2. Paragraph Meaning		1st			.26	.26	.13			
3. Vocabulary		1st			.20	.20	.10			
4. Spelling										

8. Saldate et al. (1985)	strong	2nd	2 years	immediate						
		3rd	3 years		.91	.93	.42			
						Study d = .15				
9. Rothfarb	strong	K			–.09	–.09		unacceptable	not reviewed	??
	light	1st			–.46	–.46				
						2 cohorts				
10. Murtagh (1982)	light	1st	1 year[18]	immediate	1.13	1.20		flawed	not reviewed	?
Oral		2nd	2 years	immediate	.45	.47				
		3rd	3 years	immediate	.98	1.04				
						3 cohorts (avg d = .89)				

(continued)

TABLE 3 Strong, Light, and Weak Bilingual Education Studies (continued)

					ES					
Study/Exam	Quality of BE Program	Grade Level Tested	Length of Program	Time of Study	g	d (McField, 2002)	r	Sound/Flawed Methodology?	Greene (1998)	Rossell & Baker (1996)
11. Huzar (1973)	weak	2nd	2 years	immediate	.01	.01	.01	sound	.18 Rdg.	TBE = submersion
		3rd	3 years		.31	.31	.15		.18 Eng.	
						2 cohorts				
12. Morgan (1971)	weak		7 months	immediate						
1. Word Reading		1st			.37	.38	.19	flawed	not reviewed	TBE > submersion
2. Paragraph Meaning		1st			.26	.26	.13			
3. Vocabulary		1st			.20	.20	.10			
4. Spelling		1st			.44	.44	.21			
5. Word Study Skills		1st			.23	.23	.11			
						Study d = .30				
13. Kaufman (1968)	weak									
Retest II (5/1964)			9 months	immediate				sound	.20 Rdg.	TBE = submersion
									.20 Eng.	TBE = submersion
School B	Word Meaning				.05	.05	.02			
	Paragraph Meaning				.48	.49	.24			
						Cohort d = .27				
Retest III (3/1965)			16 months	immediate						
School A	Word Meaning				.29	.30	.15			
	Paragraph Meaning				.11	.11	.06			
						Cohort d = .20				
School B	Not Tested									

(continued)

TABLE 3 Strong, Light, and Weak Bilingual Education Studies (continued)

Study/Exam	Quality of BE Program	Grade Level Tested	Length of Program	Time of Study	ES g	ES d	ES r	Sound/Flawed Methodology?	Greene (1998)	Rossell & Baker (1996)
					(McField, 2002)					
14. Covey (1973)	undefined									
Stanford		**9th**	**9 months**	**immediate**	**.74**	.74	**.35**	flawed	.74 Rdg.	
ITED/Correct & Appropriate Writing		9th	9 months		.48	.48	.23		.34 Eng.	
					Study d = .61					
15. Bacon et al. (1982)	undefined	8th (2nd–5th)	4 yrs (after 2 yrs)	graduates	.80	.82	.38	flawed	.68 Rdg.	?
		8th (1st–5th)	5 yrs		.95	.98	.44		.79 Eng.	TBE > submersion
					2 cohorts					

program studies or all accepta
consideration their different sa

Fixed and random effects m
tions as follows. For the sets of
undefined in terms of bilingual
calculated using a fixed effects
all studies included in the pres
gram quality, a random effects n
of different program quality ca
necessitates the use of a fixed
category, whereas in contrast, b
be used for calculating an overa

The effect sizes calculated along
examined for patterns within a
findings for each category, and

11 *Strong* Biling
3 Me
8 Met

4 *Light* Biling
4 Met

5 *Weak* Biling
4 Me
1 Met

3 *Undefined* Biling
3 Met

Grand Weighted Eff
and Undefi

Hypothesis 1: For studies of
search quality, the better the
light, or weak bilingual educa
outcomes (reported in effect

Finding 1: For acceptable stu
quality had higher effect sizes
of the light and undefined stu

program studies or all acceptable research quality studies) by taking into consideration their different sample sizes and effect sizes.

Fixed and random effects models were used to guide effect size calculations as follows. For the sets of studies found to be strong, light, weak and undefined in terms of bilingual program quality, summary effect sizes were calculated using a fixed effects model. For a grand average effect size for all studies included in the present review, including those of varying program quality, a random effects model was used. The assumption that studies of different program quality categories would exhibit different effect sizes necessitates the use of a fixed effects model within each program quality category, whereas in contrast, by definition, a random effects model would be used for calculating an overall grand mean across all studies.

FINDINGS

The effect sizes calculated along research quality and program quality were examined for patterns within and across each program quality type. The findings for each category, and results of hypotheses tested, were as follows.

11 *Strong* Bilingual Program Cohorts	d = .56
3 Methodologically Sound	**d = .41**
8 Methodologically Flawed	d = .58
4 *Light* Bilingual Program Cohorts	d = –.02
4 Methodologically Flawed	d = –.02
5 *Weak* Bilingual Program Cohorts	d = .24
4 Methodologically Sound	**d = .19**
1 Methodologically Flawed	d = .30
3 *Undefined* Bilingual Program Cohorts	d = .54
3 Methodologically Flawed	d = .54
Grand Weighted Effect Size for Srong, Light, Weak and Undefined Program Cohorts	
23 cohorts	d = .44

Hypothesis 1: For studies of both acceptable and unacceptable research quality, the better the bilingual education program (strong, light, or weak bilingual education programs), the better the students' outcomes (reported in effect sizes).

Finding 1: For acceptable studies only, studies with strong program quality had higher effect sizes than studies with weak programs. None of the light and undefined studies were of acceptable research quality;

thus the average combined effect size for these categories could not be calculated. For combined effect size computations for both acceptable and unacceptable research quality, strong bilingual education programs had the highest effect sizes, followed by undefined, weak, and light studies. However, it must be noted again that for light and undefined studies, none were of acceptable research quality; thus the combined effect size computations for both acceptable and unacceptable strong and weak bilingual education programs were compared to effect sizes of only unacceptable light and undefined bilingual education programs.

Hypothesis 2: For studies of both acceptable and unacceptable research quality, students in undefined bilingual education programs will demonstrate weak effect sizes relative to students in strong, light, and weak bilingual education programs.

Finding 2: All of the studies in this category were unacceptable. Thus effect sizes for acceptable undefined studies only could not be calculated. When unacceptable undefined studies were compared to unacceptable strong, light, weak, and undefined programs, the mean effect size for unacceptable undefined bilingual education programs was higher than the mean effect size for unacceptable light and weak bilingual education programs, but slightly lower than for unacceptable strong bilingual education programs. Comparisons between acceptable undefined bilingual education studies and acceptable strong, light and weak programs could not be made, since there were no acceptable undefined bilingual education program studies.

Hypothesis 3: The tighter the research quality (research design, control for bias, etc.) *and* program quality, the higher the effect size.

Finding 3: The mean effect size for studies of acceptable research quality and strong program quality was higher than the mean effect size of studies of acceptable research quality and weak program quality. Specifically, for methodologically sound studies only, the average effect sizes by varying program quality were as follows:

Strong Bilingual Program	.41
Light	n/a
Weak	.19
Undefined	n/a

This is a key suggestive finding, although the pattern of higher effect sizes for studies of higher program quality could not be fully tested due to the lack

of acceptable studies found in
There is some evidence to sugg
effect sizes by program quality a
of effect sizes based only on res
tant distinction from previous r
age effect sizes along the catego
quality only (with different stud
the set of studies with flawed (u
than the set of studies with sou
ference was nearly double for u
acceptable research designs (d =
ies only, the effect size for stron
nearly double the effect size for
Acceptable studies with light a
computed since there were no
view. It is of particular importan
the quality of the bilingual edu
research design to conduct the
greater magnitude (d = .41) tha
sidered (d = .26). Moreover, usi
allow for the effects of bilingua
amined more thoroughly and s
of effect sizes yielded by the twc
research design is an important
tant to consider program qualit
bilingual education programs,
bilingual education programs ca

Methodologically Sound Studie
(7 Cohorts from 5 studies =

Methodologically Flawed Studi
(16 Cohorts from 10 studies

It is important to note that
studies that met established cri
ception of Rolstad et al. 2005, w
able and unacceptable studies t
quality together.

Fail-Safe N Calculations

Fail-Safe N calculations were
studies of negative outcomes for

:or these categories could not
computations for both accept-
y, strong bilingual education
followed by undefined, weak,
ıoted again that for light and
able research quality; thus the
ɔoth acceptable and unaccept-
n programs were compared to
ıd undefined bilingual educa-

eptable and unacceptable re-
ɔilingual education programs
ve to students in strong, light,
3.

gory were unacceptable. Thus
udies only could not be calcu-
;tudies were compared to un-
ıdefined programs, the mean
bilingual education programs
r unacceptable light and weak
htly lower than for unaccept-
;rams. Comparisons between
ıtion studies and acceptable
not be made, since there were
cation program studies.

quality (research design, con-
the higher the effect size.

tudies of acceptable research
; higher than the mean effect
uality and weak program qual-
ɔund studies only, the average
vere as follows:

ram	.41
	n/a
	.19
	n/a

ı the pattern of higher effect sizes
. not be fully tested due to the lack

of acceptable studies found in the light and undefined program categories. There is some evidence to suggest that there may exist a different pattern of effect sizes by program quality *and* research quality, compared to the pattern of effect sizes based only on research quality. This finding reveals an important distinction from previous meta-analyses in the field. The weighted average effect sizes along the categories of acceptable and unacceptable *research quality only* (with different studies of mixed program quality) revealed that the set of studies with flawed (unacceptable) design had a higher effect size than the set of studies with sound (acceptable) design. In this study, the difference was nearly double for unacceptable designs (d = .48) over those with acceptable research designs (d = .26). Within the category of acceptable studies only, the effect size for strong bilingual education programs (d = .41) was nearly double the effect size for weak bilingual education programs (d = .19). Acceptable studies with light and undefined program quality could not be computed since there were no studies in these categories in the present review. It is of particular importance and interest to note that considering both the quality of the bilingual education program as well as the quality of the research design to conduct the calculations revealed an effect size of much greater magnitude (d = .41) than if only the research design quality was considered (d = .26). Moreover, using the more comprehensive approach would allow for the effects of bilingual education program components to be examined more thoroughly and systematically. The comparison of the pattern of effect sizes yielded by the two sets of analyses reveals that, while adequate research design is an important factor to consider, it is also critically important to consider program quality when considering the degree of impact of bilingual education programs, so that the impact of the quality or type of bilingual education programs can be measured accurately.

Methodologically Sound Studies:	
(7 Cohorts from 5 studies = 3 strong bilingual program, 2 weak)	d = .26
Methodologically Flawed Studies:	
(16 Cohorts from 10 studies = 5 strong, 1 weak, and 4 undefined)	d = .48

It is important to note that all previous meta-analyses have examined studies that met established criteria for research quality only (with the exception of Rolstad et al. 2005, which calculated effect sizes for both acceptable and unacceptable studies together) not research quality and program quality together.

Fail-Safe N Calculations

Fail-Safe N calculations were conducted in order to determine how many studies of negative outcomes for bilingual education would have to be located in order to render the findings of this review insignificant. 989 studies would be needed in order to bring the grand mean effect size for all studies of varying program and research quality, or an average d = .44 down to d = .01. The .01 was used as a benchmark with the premise that a bilingual program that produces equal or better effect sizes is effective, since both the primary language and English are used to facilitate the development of English, with outcomes similar to control group students. As an additional point of reference, 78 studies of small or negative outcomes would be needed in order to bring the average d = .44 found in this review down to d = .10.

Comparison with Previous Reviews of Bilingual Education

On the whole, bilingual education has been found to have positive outcomes, when compared to English-Only programs, with effects ranging from extremely weak to strong: (narrative or vote count reviews, listed chronologically—see Zappert and Cruz, 1977; Troike, 1978; Krashen, 1996 on Baker & de Kanter, 1983; Cziko, 1991; Lam, 1992; Krashen, 1996 on Rossell & Baker, 1996; Demmert & Towner, 2003; meta-analyses, also listed chronologically—see Okada et al. 1982; Willig, 1985; Oh, 1987; Greene, 1998; McField, 2002; Rolstad et al. 2005; Rossell & Kuder, 2005; Slavin & Cheung, 2005; Francis, Lesaux & August, 2006; McField, 2007).

According to Cohen's (1977) standard, the average effect size for bilingual education programs is moderate (between small and large, according to Cohen, 1977). According to Tallmadge (1977), the average effect size for bilingual education programs is educationally significant. In any case, the effect of bilingual education programs is positive, with about a four-month advantage (d = .41) over all-English programs for strong bilingual programs of acceptable research design.

CONCLUSIONS

Several conclusions can be drawn from the present meta-analysis. First, a review of program quality (consideration of both the definition and implementation of bilingual education programs) is equally important as is a discussion of research quality. In the present review, for studies with acceptable research designs, the average effect sizes followed the expected pattern of strong bilingual education programs showing greater efficacy (d = .41) than weak bilingual education programs (d = .19). Light programs could not be tested due to the lack of studies in this category of sound research quality. In contrast, focusing only on methodological rigor

did not bear out the expected
studies yielded higher effect siz
studies (d = .26). It is of partic
considering both the quality of
the quality of the research desi
effect size of much greater ma
design quality was considered (

Second, meta-analysis allows f
pared to narrative reviews or vo
of primary studies are involved.
popularity in high quality quanti
for ELLs needs to continue docu
using meta-analysis. The need in
methodology does not preclude
grams and effective components

Third, on the whole, the fin
findings of previous major revie
conducted to date, in that positiv
tion. The strikingly similar result
support for bilingual education
demically in English, and as a me
ly than using all-English method
doubt on claims that all-English
dated by law, as has been done in

There is no doubt that, whe
language instruction is part of
research continues to yield info
cessful programs for ELLs, it is
bilingual education in the futur

IM

1. Meta-analysis should be uti
bilingual education.
2. Clear bilingual program d
the original studies and re
analytic reviews. Studies wi
program features are not a
the field.
3. Bilingual education contin
English language developn
restrictions in the impleme

insignificant. 989 studies would
an effect size for all studies of
average d = .44 down to d = .01.
:emise that a bilingual program
ffective, since both the primary
ıe development of English, with
As an additional point of refer-
es would be needed in order to
/ down to d = .10.

ıf Bilingual

been found to have positive
y programs, with effects rang-
ive or vote count reviews, listed
77; Troike, 1978; Krashen, 1996
Lam, 1992; Krashen, 1996 on
2003; meta-analyses, also listed
'illig, 1985; Oh, 1987; Greene,
:ossell & Kuder, 2005; Slavin &
)06; McField, 2007).
:he average effect size for bilin-
veen small and large, according
(1977), the average effect size
ionally significant. In any case,
s is positive, with about a four-
: programs for strong bilingual

۱S

e present meta-analysis. First, a
f both the definition and imple-
ıs) is equally important as is a
ent review, for studies with ac-
ect sizes followed the expected
grams showing greater efficacy
programs (d = .19). Light pro-
k of studies in this category of
g only on methodological rigor

did not bear out the expected outcomes, since flawed bilingual education studies yielded higher effect sizes (d = .48) than sound bilingual education studies (d = .26). It is of particular importance and interest to note that considering both the quality of the bilingual education program as well as the quality of the research design to conduct the calculations revealed an effect size of much greater magnitude (d = .41) than if only the research design quality was considered (d = .26).

Second, meta-analysis allows for a clearer summary of the field when compared to narrative reviews or vote counts, especially when a sizable number of primary studies are involved. Given that effect sizes have gained greater popularity in high quality quantitative research studies, the field of programs for ELLs needs to continue documenting, analyzing and reviewing programs using meta-analysis. The need in the field for such a consistent quantitative methodology does not preclude the need to describe and document programs and effective components therein using qualitative methods.

Third, on the whole, the findings of this review are consistent with the findings of previous major reviews, including all major quantitative reviews conducted to date, in that positive outcomes were found for bilingual education. The strikingly similar results from different meta-analyses provide clear support for bilingual education as a means of helping children succeed academically in English, and as a means for acquiring English much more rapidly than using all-English methods and programs. The results also cast strong doubt on claims that all-English approaches are superior and should be mandated by law, as has been done in California, Arizona, and Massachusetts.

There is no doubt that, when it comes to English acquisition, native-language instruction is part of the solution, not part of the problem. As research continues to yield information about the factors that predict successful programs for ELLs, it is likely that we will see larger effect sizes for bilingual education in the future.

IMPLICATIONS

1. Meta-analysis should be utilized to periodically review the field of bilingual education.
2. Clear bilingual program descriptions need to be included both in the original studies and reviews, to facilitate analysis and use in meta-analytic reviews. Studies with unclear descriptions of instruction and program features are not acceptable as they do little to illuminate the field.
3. Bilingual education continues to demonstrate strength in providing English language development for ELLs. There is no need for strict restrictions in the implementation of these programs. Popular ideol-

ogy often overshadows the efficacy and power of bilingual education programs, but the present review is one among many that suggests that popular ideology and corresponding English-only language policies need to be systematically questioned, reexamined, and overhauled, rather than a uniform program mandated regardless of research base, context (e.g., local needs), and resources. The findings of this review strongly suggest that local educational agencies ought to be given the flexibility to choose the best language program for students, with input from all appropriate stakeholders, including parents, teachers, educational leaders, and the students themselves.

FUTURE DIRECTIONS

1. The field is beginning to settle on a metric, as noted above about the use of different statistics for effect size calculations. In light of the advances in statistical considerations and the incorporation of Hedges' adjusted g in the two most recent meta-analyses (Francis Lesaux, & August, 2006; Slavin & Cheung, 2005), all future meta-analyses should be explicit and clear about the use of different effect size metrics and the differential impacts therein.
2. All meta-analyses on programs for ELLs need to consider random vs. fixed effects in effect size calculations. As evidenced in the present review, analyses and reporting of different sets and subsets of bilingual studies can look very different. Using grounded theory (e.g., the presence or absence of key program quality components) to drive statistical analysis, random vs. fixed effects models need to be explored, and used correspondingly and appropriately. The present study may be used as a guide to inform the use of fixed vs. random effects in considering the impact of programs for ELLs.
3. The findings of the present study ought to be extended using additional primary studies of bilingual education and English-only programs. The field of programs for ELLs has made significant advances over the past two decades, and current primary studies ought to be analyzed for research design and program quality components in order to test the relative efficacy of strong, light, weak and undefined bilingual education programs.

NOTES

1. This federal study was subsequently published by Lawrence Erlbaum in 2006.

2. The effect sizes are for all
Slavin and Cheung (2005), w
reviewers included only stud
treatments or in which other
Mahoney, and Glass (2005)
Rossell and Kuder (2005)
studies covered in Slavin an
Spanish-speaking children i
culated an average effect siz
measure, compared to Gree
lations for most individual s
calculated an effect size of –
not use the final year of the s
year, based on Rossell's regr
Using a sample expanded b
the test but who did not take
in Rossell, p. 91, Table 4.6).
3. McField (2002) considered
than one effect size in some
and Glass, 2005) are not incl
and Kuder (2005) note that
tion. In Lindholm (1991), th
no significant difference betw
3 but it was impossible to com
The Medrano (1986) effect
(1988) for grade 3 results.
4. Maldonado (1994)–Given the
that something could have le
gains in their posttest scores
This would give us our very l
value does not match an ES
more reasonable, yet large ES
more reasonable to use the
Panel, 2006; Slavin & Cheung
the assumption that the num
perimental group were trans
but standard errors. While thi
methodologically sound to use
ES instead. Doing so results in
Rossell and Kuder consider
the effect size is "unbelievable
size could have been due at lea
er assigned to the treatment g
bilingual special education' an
abilities. The control group te
with bilingual students with l
used by the experimental grou

ɔower of bilingual education
among many that suggests
g English-only language
ɔned, reexamined, and
ram mandated regardless of
), and resources. The find-
ocal educational agencies
se the best language program
riate stakeholders, including
nd the students themselves.

ɔNS

ric, as noted above about
ze calculations. In light of
s and the incorporation of
nt meta-analyses (Francis
ıg, 2005), all future meta-
out the use of different effect
s therein.
need to consider random
s. As evidenced in the pres-
ferent sets and subsets of
. Using grounded theory
ogram quality components)
xed effects models need
gly and appropriately. The
inform the use of fixed vs.
ct of programs for ELLs.
t to be extended using ad-
ucation and English-only pro-
has made significant advances
primary studies ought to be
am quality components in
ng, light, weak and undefined

ed by Lawrence Erlbaum in 2006.

2. The effect sizes are for all measures of achievement combined, except for Slavin and Cheung (2005), who considered only tests of English reading. Most reviewers included only studies in which students were randomly assigned to treatments or in which other means of matching students were used. Rolstad, Mahoney, and Glass (2005) did not feature this requirement.

   Rossell and Kuder (2005) arrived at an average effect size of .14 for the studies covered in Slavin and Cheung, limiting their analysis to studies of Spanish-speaking children in elementary school (14 studies). They also calculated an average effect size of –.07 for Greene's studies using reading as a measure, compared to Greene's result of .21 for reading. Effect size calculations for most individual studies were very similar, but Rossell and Kudar calculated an effect size of –.25 for Rossell (1990), claiming that Greene did not use the final year of the study. We estimated an effect of size of .10 for that year, based on Rossell's regression results (from Rossell, 1990, appendix 2). Using a sample expanded by adding chance scores for students eligible for the test but who did not take it, the effect size moves to a negative 1.66 (data in Rossell, p. 91, Table 4.6).
3. McField (2002) considered separate cohorts; hence the presence of more than one effect size in some cases. Gersten's studies (from Rolstad, Mahoney, and Glass, 2005) are not included; for discussion, see Krashen (1996). Rossell and Kuder (2005) note that Gersten (1985) did not involve bilingual education. In Lindholm (1991), the effect size was based only on grade 2; there was no significant difference between bilingual and comparison students in grade 3 but it was impossible to compute effect sizes from the information provided. The Medrano (1986) effect size is based on grade 6 results. See Medrano (1988) for grade 3 results.
4. Maldonado (1994)–Given the population and the controls used it is possible that something could have led the control group to shut down and not show gains in their posttest scores while the experimental group achieved gains. This would give us our very large ES of 7. However, given that the t statistic value does not match an ES of 7, and since 7 is very large compared to the more reasonable, yet large ES of 1.73 derived from the t value, then it seems more reasonable to use the 1.73 value for our ES. Others (National Literacy Panel, 2006; Slavin & Cheung, 2005) have reported an ES of 2.25, based on the assumption that the numbers for the pre and post test scores for the experimental group were transposed, and that the SD as stated were not SDs but standard errors. While this assumption seems reasonable, it seems more methodologically sound to use the t value given by the author to calculate an ES instead. Doing so results in an ES of 1.73.

   Rossell and Kuder consider Maldonado (1994) to be an "outlier" because the effect size is "unbelievable." They note that the exceptionally large effect size could have been due at least in part to teacher differences: "[T]he teacher assigned to the treatment group had experience working with 'integrated bilingual special education' and teaching bilingual students with learning disabilities. The control group teacher apparently had no experience working with bilingual students with learning disabilities... The teaching strategies used by the experimental group teacher [also] include a wide range of strategies beyond the language of instruction" (p. 56). In addition, the gains made by the experimental group were so "astonishing" that Rossell and Kuder say that "one can only wonder if the researcher made a mathematical or other kind of error" (p. 59).

5. Cottrell (1971)—Only the results for first grade students were calculated. Calculations were not done for the cohort of kindergarteners' scores, due to the fact that kindergarteners were tested on readiness measures for both pretests and posttests, and it was unclear whether reading comprehension skills could be detected by these measures.
6. Franks (1988)—The large differences in pre-test scores between the experimental and the control groups could be a cause for concern, especially if the control group had scored lower than the experimental group. However, since the control group outscored the experimental group the possibility of scores being influenced by a ceiling effect can be eliminated. Furthermore, it implies that the gain scores would have likely been higher if the low pre-test scores for the experimental group had been adjusted for. This means that by using the scores "as is," the effect size presented here is an underestimate of this study's true effect size.

   Furthermore, the SD of the experimental group at the pre-test levels was very different from those of the control pre-test scores. This was cause for concern. However, because the experimental group post-test SD was similar to the control group's SD the two groups do appear to be similar, but the large pre-test SD of the experimental group could be due to the fact that the pre-test scores of the experimental group were much lower than those of the control group. However, some members in the experimental group may have scored as high, or higher, and other lower than the experimental group before treatment. This may explain some of the discrepancy between the two SDs. Once the experimental group gained as much, and later, more than the control group, their scores "settled" around the mean more like the control group scores. By pooling the pre-test experimental SD with the other SDs, we have created a larger SD and made the ES estimate more conservative.

   The pooling of the pre-test experimental SD and the lack of control of differences for the large pre-test scores, makes our ES calculation very conservative.
7. Same considerations as noted above in Table 1. To restate, McField (2002) considered separate cohorts, hence the presence of more than one effect size in some cases. Gersten's studies (from Rolstad, Mahoney, and Glass, 2005) are not included; for discussion, see Krashen (1996). Rossell and Kuder (2005) note that Gersten (1985) did not involve bilingual education. In Lindholm (1991), the effect size was based only on grade 2; there was no significant difference between bilingual and comparison students in grade 3 but it was impossible to compute effect sizes from the information provided. The Medrano (1986) effect size is based on grade 6 results. See Medrano (1988) for grade 3 results.

   Rossell and Kuder consider Maldonado (1994) to be an "outlier" because the effect size is "unbelievable." They note that the exceptionally large effect size could have been due at least in part to teacher differences: "[T]he teacher assigned to the treatment group had experience working with 'integrated

bilingual special education'
abilities. The control grou
with bilingual students wit
used by the experimental g
gies beyond the language o
by the experimental group
that "one can only wonder
kind of error" (p. 59).

8. In the present meta-analysis
in the primary study, or if th
reported test scores for mult
reported as a study d herein
all years of treatment should
9. Maldonado (1994)–Given t
that something could have
gains in their posttest scor
This would give us our very
value does not match an E
more reasonable, yet large
more reasonable to use the
Panel, 2006; Slavin & Cheu
the assumption that the nu
perimental group were tra
but standard errors. While
methodologically sound to
ES instead. Doing so results
10. This study was found to be
moved prior to analyses in
in Table 3. However, in McF
average d for undefined pro
of undefined and unaccept
11. Effect sizes were calculated
not available.
12. Cottrell (1971)–Only the re
culations were not done for
fact that kindergarteners we
and posttests, and it was unc
be detected by these measu
13. Franks (1988)–The large di
mental and the control gro
the control group had scor
since the control group ou
scores being influenced by a
implies that the gain scores
scores for the experimental
using the scores "as is," the
this study's true effect size.

bilingual special education' and teaching bilingual students with learning disabilities. The control group teacher apparently had no experience working with bilingual students with learning disabilities... The teaching strategies used by the experimental group teacher [also] include a wide range of strategies beyond the language of instruction" (p. 56). In addition, the gains made by the experimental group were so "astonishing" that Rossell and Kuder say that "one can only wonder if the researcher made a mathematical or other kind of error" (p. 59).

8. In the present meta-analysis, if more than one type of test score was reported in the primary study, or if the primary study utilized a longitudinal design and reported test scores for multiple years, an average effect size was calculated and reported as a study d herein. The rationale was that all types of test scores and all years of treatment should be considered to capture an average effect size.
9. Maldonado (1994)–Given the population and the controls used it is possible that something could have led the control group to shut down and not show gains in their posttest scores while the experimental group achieved gains. This would give us our very large ES of 7. However, given that the t statistic value does not match an ES of 7, and since 7 is very large compared to the more reasonable, yet large ES of 1.73 derived from the t value, then it seems more reasonable to use the 1.73 value for our ES. Others (National Literacy Panel, 2006; Slavin & Cheung, 2005) have reported an ES of 2.25, based on the assumption that the numbers for the pre and post test scores for the experimental group were transposed, and that the SD as stated were not SDs but standard errors. While this assumption seems reasonable, it seems more methodologically sound to use the t value given by the author to calculate an ES instead. Doing so results in an ES of 1.73.
10. This study was found to be an outlier in the test of homogeneity, and removed prior to analyses in this present analysis. Thus, the study is not listed in Table 3. However, in McField (2002), it was left in for the calculation of the average d for undefined programs, since that category was comprised entirely of undefined and unacceptable studies.
11. Effect sizes were calculated using unadjusted means, as other statistics were not available.
12. Cottrell (1971)–Only the results for first grade students were calculated. Calculations were not done for the cohort of kindergarteners' scores, due to the fact that kindergarteners were tested on readiness measures for both pretests and posttests, and it was unclear whether reading comprehension skills could be detected by these measures.
13. Franks (1988)–The large differences in pre-test scores between the experimental and the control groups could be a cause for concern, especially if the control group had scored lower than the experimental group. However, since the control group outscored the experimental group the possibility of scores being influenced by a ceiling effect can be eliminated. Furthermore, it implies that the gain scores would have likely been higher if the low pre-test scores for the experimental group had been adjusted for. This means that by using the scores "as is," the effect size presented here is an underestimate of this study's true effect size.

Furthermore, the SD of the experimental group at the pre-test levels was very different from those of the control pre-test scores. This was cause for concern. However, because the experimental group post-test SD was similar to the control group's SD the two groups do appear to be similar, but the large pre-test SD of the experimental group could be due to the fact that the pre-test scores of the experimental group were much lower than those of the control group. However, some members in the experimental group may have scored as high, or higher, and other lower than the experimental group before treatment. This may explain some of the discrepancy between the two SDs. Once the experimental group gained as much, and later, more than the control group, their scores "settled" around the mean more like the control group scores. By pooling the pre-test experimental SD with the other SDs, we have created a larger SD and made the ES estimate more conservative.

The pooling of the pre-test experimental SD and the lack of control of differences for the large pre-test scores, makes our ES calculation very conservative.

14. In Doebler & Mardis (1980), all students, both experimental and control, were given Choctaw instruction in Kindergarten, and ESL (CORE English) in 1st grade. Then in the 3rd grade, experimental students were given strong BE and control group students were taught using mainstream English.
15. Reading denotes scores for reading comprehension in English.
16. English denotes scores for Language Arts such as mechanics and skills.
17. In Franks (1988), pretests were administered after the treatment was in effect.
18. In Murtagh (1982), a study from Australia, all students were tested at the beginning of the academic year after summer vacation. For example, 1st graders had participated in the program in preschool for one academic year, and were tested at the beginning of 1st grade. It is unclear whether, as is the case in some parts of Australia, preschool was used synonymously with what is referred to as kindergarten or first year in an elementary school setting in the U.S.; or whether preschool referred to a broader and longer program (for instance, over four terms before starting formal schooling). In any case, testing was done after summer break; thus the program effects were probably a conservative measure.

REFERENCES

Methods Literature: Meta-Analysis and General

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (Rev. ed.). New York: Academic Press.

Grissom, R. J., & Kim, J. J. (2005). *Effect Sizes for Research: A Broad Practical Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Light, R. J., & Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press.

Wolf, F. M. (1986). *Meta-Analysis: Quantitative Methods for Research Synthesis*. Beverly Hills, CA: Sage Publications.

Glass, G. V. (1976). Primary, sec
Researcher, 5, 3–8.

Glass, G. V., McGaw, B., Smith, M
Hills, CA: Sage Publication

Hedges, L. V. (1982). Estimation
ments. *Psychological Bulletin*

Hedges, L. V., & Olkin, I. (1985).
Academy Press.

Rosenthal, R. (1984). *Meta-analys*

Rosenthal, R. (1995). Writing m
183–192.

Rosenthal, R., & Rosnow, R. L. (
data analysis, (2nd ed.) Bost

Rosenthal, R., & Rubin, D. (1982
Psychological Bulletin, 92, 50

Tallmadge, G. K. (1977). *The Joi*
DC: National Institute of Ed

Meta-Analyses of Bilingu
and Related Articles

Francis, D.J., Lesaux, N., & Augus
gust & T. Shanahan (Eds), *L*
of the national literacy panel
413). Hillsdale, NJ: Lawrenc

Greene, J. (1999). A meta-analysis
cation research. *Bilingual Re*

Greene, J. (1998). *A meta-analysis o*
CA: Tomas Rivera Policy Inst

Krashen, S., & McField, G. (2005)
latest evidence. *Language Le*

McField, G. (2002). *Does program qu*
tion studies. Ph.D. Dissertatio

McField, G. (2007). *The role of progr*
meta-analyses. Report funded
Research Association and Ins
of Education.

Oh, S. S. (1987). *A comparative stud*
bilingual education programs f
doctoral dissertation. The Fl

Okada, M., Besel, R., Glass, G. V., M
sis of reported evaluation and re
tion: Basic projects, final report.
Bilingual Research.

group at the pre-test levels was
-test scores. This was cause for
l group post-test SD was similar
o appear to be similar, but the
could be due to the fact that
were much lower than those of
in the experimental group may
er than the experimental group
he discrepancy between the two
much, and later, more than the
the mean more like the control
ental SD with the other SDs, we
imate more conservative.
and the lack of control of differ-
ES calculation very conservative.
oth experimental and control,
en, and ESL (CORE English) in
al students were given strong BE
mainstream English.
ension in English.
h as mechanics and skills.
after the treatment was in effect.
all students were tested at the
vacation. For example, 1st grad-
hool for one academic year, and
s unclear whether, as is the case
d synonymously with what is re-
lementary school setting in the
oader and longer program (for
al schooling). In any case, test-
rogram effects were probably a

General

havioral Sciences (Rev. ed.). New

earch: A Broad Practical Approach.

The Science of Reviewing Research.

ods for Research Synthesis. Beverly

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Glass, G. V., McGaw, B., Smith, M. L. (1981). *Meta-Analysis in Social Research.* Beverly Hills, CA: Sage Publications.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academy Press.

Rosenthal, R. (1984). *Meta-analysis for social research.* Beverly Hills, CA: Sage.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118*(2), 183–192.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis,* (2nd ed.) Boston, MA: McGraw Hill.

Rosenthal, R., & Rubin, D. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin,* 92, 500–504.

Tallmadge, G. K. (1977). *The Joint dissemination review panel ideabook.* Washington, DC: National Institute of Education and U.S. Office of Education.

Meta-Analyses of Bilingual Programs and Related Articles

Francis, D.J., Lesaux, N., & August, D. (2006). Language of instruction. In D. August & T. Shanahan (Eds), *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth* (pp. 365–413). Hillsdale, NJ: Lawrence Erlbaum Associates.

Greene, J. (1999). A meta-analysis of the Rossell and Baker review of bilingual education research. *Bilingual Research Journal,* April, 1999.

Greene, J. (1998). *A meta-analysis of the effectiveness of bilingual education.* Claremont, CA: Tomas Rivera Policy Institute.

Krashen, S., & McField, G. (2005) What works for English Learners? Reviewing the latest evidence. *Language Learner, 29*(3), 7–11.

McField, G. (2002). *Does program quality matter? A meta-analysis of select bilingual education studies.* Ph.D. Dissertation, University of Southern California.

McField, G. (2007). *The role of program quality and research quality in bilingual education meta-analyses.* Report funded by a grant award from the American Educational Research Association and Institute for Education Sciences, U.S. Department of Education.

Oh, S. S. (1987). *A comparative study of quantitative vs. qualitative synthesis of Title VII bilingual education programs for Asian children in New York City.* Unpublished doctoral dissertation. The Florida State University.

Okada, M., Besel, R., Glass, G. V., Montoya-Tannatt, L., & Bachelor, P. (1982). *Synthesis of reported evaluation and research evidence on the effectiveness of bilingual education: Basic projects, final report: Tasks 1–6.* Los Alamitos, CA: National Center for Bilingual Research.

Rolstad, K., Mahoney, K., and Glass, G. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy 19*(4): 572–594.

Rossell, C. and Kuder, J. (2005). Meta-murky: A rebuttal to recent meta-analyses of bilingual education. In J. Sohn (Ed.) *The effectiveness of bilingual school programs for immigrant children.* Berlin: Arbeitsstelle Interkulturelle Konflikte und gesellschaftliche Integration (AKI). pp. 43–76.

Slavin, R. and Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research 75*(2): 247–284.

Willig, A. (1987). Examining bilingual education research through meta-analysis and narrative review: A response to Baker. *Review of Educational Research, 57*(3), 363–376.

Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55*(3), 269–317.

Non-Meta-Analytic Cumulative Reviews of Bilingual Programs

Baker, K. A., & De Kanter, A. A. (1981). *Effectiveness of bilingual education: A review of the literature.* Washington, D.C.: U.S. Department of Education, Office of Planning, Budget and Evaluation.

Baker, K. A., & De Kanter, A. A. (1983). Federal policy and the effectiveness of bilingual education. In K.A. Baker and A. A. De Kanter (eds.), *Bilingual education: A reappraisal of federal policy* (pp. 33–86). Lexington, MA: Lexington Books.

Cziko, G. A. (1992). The evaluation of bilingual education: From necessity and probability to possibility. *Educational Researcher, 21*(2), 10–15.

Cummins (2000) Theory in bilingual education research & review of research. In Ovando & McLaren (Eds.), *The politics of multiculturalism and bilingualeEducation: Students and teachers caught in the crossfire* (pp. 126–147). New York: McGraw Hill.

Demmert, W. G., & Towner, J.C. (2003). *A review of the research literature on the influences of culturally based education on the academic performance of Native American students.* Portland, OR: Northwest Regional Educational Laboratory.

Lam, T. C. M. (1992). Review of practices and problems in the evaluation of bilingual education. *Review of Educational Research, 62*(2), 181–203.

Rossell, C. H., & Baker, D. (1996). The educational effectiveness of bilingual education. *Research in the Teaching of English, 30*(1), 7–74.

Troike, R. C. (1978). Research evidence for the effectiveness of bilingual education. *Journal of the National Association for Bilingual Education, 3*(1), 13–24.

Zappert, L. T., & Cruz, B. R. (1977). *Bilingual education: An appraisal of empirical research.* Berkeley: Bay Area Bilingual Education League/Lau Center, Berkeley Unified School District.

General

Cummins, J. (1983). *Heritage lang*
of Education.

Cummins, J. (1999). Alternative
have a place? *Educational Re*

Krashen, S. (1996). *Under attack: T*
Language Education Associ

: big picture: A meta-analysis
anguage learners. *Educational*

ːtal to recent meta-analyses of
eness of bilingual school programs
rkulturelle Konflikte und ge-

ch on language of reading in-
of Educational Research 75(2):

search through meta-analysis
eview of Educational Research,

ı the effectiveness of bilingual
269–317.

bilingual education: A review of
t of Education, Office of Plan-

and the effectiveness of bilin-
ter (eds.), *Bilingual education:*
ton, MA: Lexington Books.
ucation: From necessity and
21(2), 10–15.
arch & review of research. In
ulturalism and bilingualeEduca-
pp. 126–147). New York: Mc-

e research literature on the influ-
performance of Native American
cational Laboratory.
ms in the evaluation of bilin-
2(2), 181–203.
ectiveness of bilingual educa-
-74.
veness of bilingual education.
ucation, 3(1), 13–24.
on: An appraisal of empirical re-
League/Lau Center, Berkeley

General

Cummins, J. (1983). *Heritage language education: A literature review.* Toronto: Ministry of Education.

Cummins, J. (1999). Alternative paradigms in bilingual education: Does theory have a place? *Educational Researcher, 28,* 26–32, 41.

Krashen, S. (1996). *Under attack: The case against bilingual education.* Culver City, CA: Language Education Associates (Distributed by Alta Book Co.).