Evaluations of Bilingual and Mother-Tongue Programs: Measures of Success and Means of Measurement

Cynthia Groff

University of Pennsylvania

This exploration of recent research studies evaluating mother-tongue programs provides a mapping of studies from selected journals in terms of both research designs and measures of success. The 669 articles published in the past five years in six selected journals were scanned, and 41 relevant impact studies were identified. These included studies addressing the impact of educational programs that use learners' mother tongue, or first language, for any portion of instructional time, whether labeled as bilingual, dual-language, or immersion, whether using transitional, maintenance, or enrichment models. Only 15% of the studies identified were quasi-experimental, and no true experimental studies were found. However, many of the studies included multiple indicators of program success. By charting the research designs of recent studies, I highlight the importance of designing research to meet evidence standards. At the same time, outlining the various outcome measures used in recent research stresses the importance of multiple means for defining and evaluating success.

Introduction

Is bilingual education helpful or harmful? Do bilingual and mother-tongue programs work? How successful are they? By whose measure of success? Many questions could be raised surrounding the issue of media of instruction. Some people feel strongly that schooling should be conducted only in the majority or standard language of a region. Others feel strongly that children deserve the opportunity to develop academic skills on the foundation of their first language. Policy-makers listen to

Working Papers in Educational Linguistics 20/2: 19-39, 2005

the debates and ask for some evidence about the success of mother-tongue and bilingual programs.

Evaluation of the success of such programs becomes important as decisions are made about the continuation and further promotion of mother-tongue education. The challenge is to know with certainty which programs work and for whom they work. From a quantitative perspective, sound research designs are required in order to be certain of results. Although a rich store of information is being developed through case studies and ethnographies, limitations in the generalizability of such studies raise doubt about their usefulness as a basis for policy decisions. Among others, the National Academy of Sciences Committee on Principles of Education Research maintains that randomized controlled trials are essential in making causal inferences and that they are superior to quasi-experiments in that they reduce the chances of error (Shavelson & Towne 2002, Mosteller & Boruch 2002). This is the view increasingly promoted in current U.S. federal policies. For example, a U.S. Department of Education policy that recently took effect gives preference to research using randomized trials and quasi-experimental designs (Glenn 2005).

On the other hand, from a qualitative perspective, the hidden assumptions in the above questions about program success may be challenged with questions about what success looks like and what is meant by whether a program "works." Different definitions of and measures of success may lead to different evaluations of the value of particular programs. Most often policy-makers are looking for numbers and scores as measures of success: signs of achievement on standardized tests. But certainly increasing scores on achievement tests is not the only goal of schools and of programs, nor could we consider it the only measure of success. What other measures of success must be considered in evaluating the impact of a program? In identifying relevant outcome variables for impact evaluations, prior research provides a useful reference. By considering the measures of success used in previous research, evaluators will not only avoid overlooking important outcomes, but will also learn how various outcome variables have been previously defined and measured (Rossi, Lipsey & Freeman 2004).

This exploration of recent research studies evaluating mother-tongue programs is organized around the tension between quantitative, experimental perspectives that encourage the use of certain research designs and qualitative perspectives that highlight variety in measures of success. The focus here is on studies addressing the impact of educational programs that use the mother tongue, or first language, of language-minority students for at least some portion of instructional time. As we chart the research designs of recent evaluation studies, the more qualitatively minded among us may reflect on the importance of can evidence standards. Through consideration of various outcome measures utilized in recent research, the more quantitatively minded among us may also begin exploring new means of analyzing success.

Meta-analyses of Program Evaluations

The theory supporting the importance of mother-tongue instruction for cognitive development during formative years is hard to ignore (see Cummins & Swain 1987, Cummins 1991), not to mention cognitive benefits of multilingualism and social and psychological benefits of language empowerment. However, strong quantitative evidence for the success of bilingual and mother-tongue programs has been minimal. Several meta-analyses have attempted to compile results of those studies that have been conducted to assess the effectiveness of using the mother tongue in instruction. Many of these studies have been conducted in the American context where English language learners are instructed with or without some use of their heritage language. Conclusions have varied, as have reviewers' definitions of acceptable studies and researchers' definitions of success.

Rossell and Baker (1996) reviewed 75 "methodologically acceptable" studies. These were defined as studies that used control-group comparison, controlled for differences between treatment and control groups either statistically or by random assignment, based results on English standardized test scores, and used appropriate statistical tests to determine differences between groups. They concluded that use of the native language in education was not beneficial for students with limited English proficiency. Later Greene (2004) conducted a meta-analysis of the Rossell and Baker literature review, narrowing down the number to 11 "methodologically acceptable" studies, using different interpretations of the same criteria. Of these, five were studies with random assignment experimental designs. From his meta-analysis Greene concludes that use of the native language is moderately beneficial for students with limited English proficiency.

A similar conclusion was reached by Willig (1985) in her meta-analysis of a literature review by Baker and de Kanter (1981). Baker and de Kanter had concluded in their literature review that in terms of their effect on English test scores, bilingual education programs were found to have negative results more often than positive results. In Willig's meta-analysis of this review, the studies were weighted according to the quality of the research designs, resulting in a positive conclusion for the effectiveness of bilingual education programs.

Greene (2004) expresses reasons for caution in drawing conclusions about the effectiveness of language instruction using these meta-analyses. First, the studies chosen are not necessarily representative of all of those available, thus results could be either more negative or more positive than estimated. In addition, the age of the studies raises some

external validity concerns as eight of the eleven studies in Greene's analysis were conducted before 1983, and these may not be representative of current techniques in mother-tongue programs. Finally, the limited sample size in some of the studies in Greene's analysis, as well as the limited number of studies, points to the need to examine more recent studies with larger sample sizes.

The What Works Clearinghouse[1] is currently conducting a similar review of research evaluations of programs for elementary school English language learners. While the present review is narrower in terms of years and journal sources, it is broader than the forthcoming What Works Clearinghouse review, which focuses only on quantitative evaluations of programs for elementary school students whose second language is English. In contrast, I include here both quantitative and qualitative studies, reviewing all of the articles that address the impact of the use of the mother tongue in an educational program.

The meta-analyses mentioned above focus on studies measuring success through standardized test scores. Evaluation of the impact of a program, however, involves more than simply measuring scores. The importance of considering multiple measures of success for mother-tongue programs is not a new concern. Many voices have spoken out questioning the validity of certain measures of achievement for English language learners (Thompson et al. 2002) and raising awareness of the inherent difficulties in assessing bilingual students' success as well as the "need for broader and fairer assessment strategies for bilingual students" (Torres-Guzmán et al. 2002). Strategies have been explored for accommodating English language learners in high-stakes assessment (Holmes & Duron 2000). Beyond the need for better assessments, many factors besides academic achievement must be considered in evaluating a program's success, including stakeholders' perceptions, attitudes and identities.

The present review does not attempt to replicate the comprehensive search or quantitative meta-analysis of previous studies. Instead, the aim is to provide a mapping of recent program evaluation studies in selected journals with regard to both research design and measures of success. The following questions will be addressed concerning recent research in the selected journals:

1) What research designs have been used in evaluating or exploring the pros and cons of mother-tongue programs?
2) What outcome variables have been used as indicators of the success of a mother-tongue program?

[1] http://www.w-w-c.org

Methods

Providing several definitions before presenting the methods used in this review is important, particularly when we consider the great variety of bilingual education and mother-tongue programs. These programs are also interpreted and misinterpreted in a variety of ways. For the purpose of this review, I will use the term "mother-tongue program" to refer to any educational program in which learners' first language is used as a medium of instruction for at least some portion of instructional time. Mother-tongue programs are contrasted with programs providing education exclusively in the learners' second language, usually the majority or standard language of the region. Bilingual education programs, which by definition should use two languages, both the first and second language of the learners, naturally fall under this definition of mother-tongue programs. While differentiation between the various program types and their bases on transitional, maintenance, or enrichment models of bilingual education (Hornberger 1991) is beyond the scope of this paper, the differences in program designs should not be ignored in comparisons of program results. My focus on methods of evaluation allows for a compilation of studies of various program designs and durations, an approach that would be problematic in meta-analyses comparing the success of programs. Evaluations here refer to assessments of program impact or outcomes, studies in which programs are examined or appraised in terms of their value or worth.

This review includes articles evaluating or exploring the pros or cons of programs whose primary characteristic is the use of the mother tongue in instruction. Although a number of evaluations in the selected volumes examined the impact of policy changes, such as the transition away from bilingual education brought about by Proposition 227 in California, such policy evaluations are excluded from this review. Also excluded are the multiple studies examining the skill development of bilinguals since they do not evaluate the effectiveness of specific educational programs.

A careful hand search of selected journals was conducted to gather articles meeting the above criteria. A hand search was used in order to provide comprehensive results and avoid the limitations of electronic searches. Several comparisons of the number of articles yielded in hand searches as compared with electronic searches have been conducted, particularly in searches for medical trials (e.g., McDonald et al. 2002). In education, Turner et al. (2003) report the results of a pilot study in which only 33% of the articles found through a hand search were found using ERIC, PsychINFO, and Sociological Abstracts. Similarly Leow and Boruch (2000) report that of 35 articles found in a hand search, only 29% were found using PsychINFO and only one of the 35 articles (3%) was found using ERIC. For this reason, I did not rely on electronic searches in identifying articles for this review.

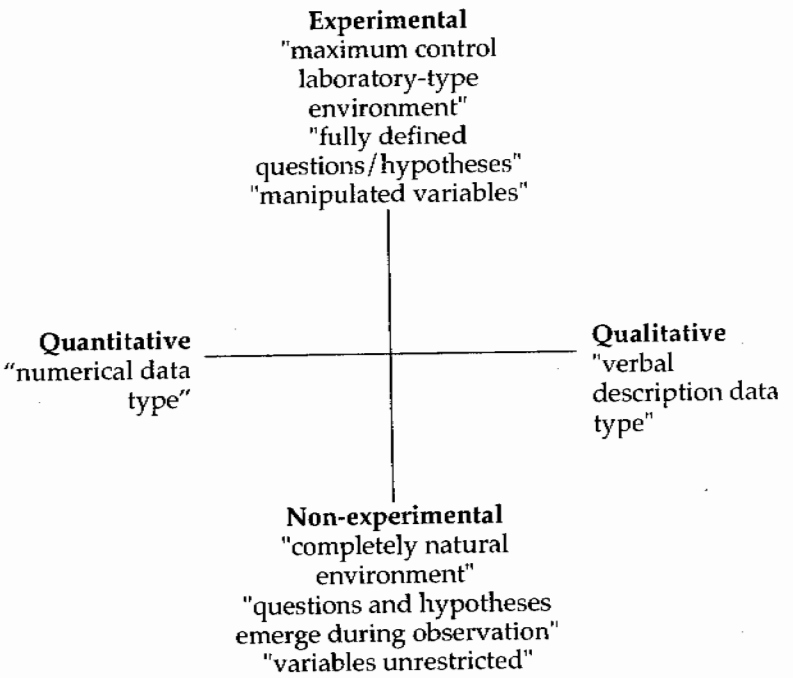
The scope of this review is narrowed in terms of time and sources for the sake of efficiency. Articles published within the past five years, from 2000 to 2004, were reviewed. Six relevant journals dealing with language and education issues were selected as sources for the articles. These were the *Bilingual Research Journal, Education Policy Analysis Archives, International Journal of Bilingual Education and Bilingualism, Journal of American Indian Education, Journal of Multilingual and Multicultural Development* and the *NABE Journal of Research and Practice.*

Articles evaluating or exploring the pros or cons of mother-tongue programs were identified and organized in terms of research design through examination of the methods used for each study. Research designs can be classified in several ways. Distinction may first be made between qualitative and quantitative studies, verbal and numerical, although this dichotomy is not always distinguishable, and many studies were identified as including both qualitative and quantitative components. A similar but separate distinction is that between experimental and non-experimental studies, those using manipulated conditions versus natural environments. These two separate classifications may be thought of as continua along which different studies may be categorized (Drew, Hardman & Hart 1985) as demonstrated in Figure 1.

The research designs were also classified as pre-experimental studies, quasi-experimental studies, and true experimental studies, along with their sub-categories as outlined by Campbell and Stanley (1966). They designate as pre-experimental all studies that examine an individual case without a controlled comparison group. In this review, descriptive and survey studies could also fit this broad criterion and be classified as pre-experimental studies. However, most of these studies are "non-experimental" in that variables are not manipulated and the environment is not controlled for the study. Many of the studies are in fact post-hoc evaluations of program impact and do not use pre-planned research designs. For the sake of comparison the studies are classified according to the various research designs, but I use the term non-experimental to refer to the studies traditionally considered pre-experimental. Classifying studies as non-experimental in this artificial model is not meant to imply inferiority or underdevelopment of particular research methods.

Figure 2 provides a list of the most common research designs relevant to program evaluation, as well as a visual representation of the research designs. All designs have a treatment (X), which for this review represents participation in a mother-tongue program, and at least one observation (O), which represents the appraisal of an outcome variable, whether the appraisal be quantitative or qualitative. Pre-experimental and here non-experimental designs examine one group without controlled comparison groups. The true experimental designs include random assignment (R) to treatment and control groups. Quasi-experi-

Figure 1
Study Types
(adapted from Drew, Hardman & Hart 1985)



mental designs use pre-tests or repeated measures as an alternative means to control for threats to validity. Campbell and Stanley (1966) list eight threats to internal validity and four threats to external validity, plotting the sources of invalidity for each design. I include their assessment of the validity potential of the designs in Figure 2 by including the number of threats to internal (In) and external (Ex) validity that are controlled by the design.

The threats to internal and external validity controlled by each research design provide some indication of the confidence placed in results using each design. Studies using randomized trials are considered most trustworthy. The results of randomized trials and non-randomized trials evaluating the same program have in many cases been found to be quite different, highlighting the need for caution in interpreting results of non-randomized studies (e.g., Boruch 1975; Duflo & Kremer 2005; Wilde & Hollister 2002).

Following the above design categories, the studies with potential for meeting specified evidence standards[2] could be identified. The What

[2] http://www.w-w-c.org/reviewprocess/standards.html

Works Clearinghouse evidence standards were used since they are representative of current guidelines for identifying scientific evidence in education research. Since the What Works Clearinghouse is currently reviewing studies of interventions for English language learners, studies that meet standards could be identified as likely to appear in their meta-analysis. In order to fully meet these evidence standards, studies must be randomized controlled trials or regression discontinuity studies having no problems with randomization, attrition, or disruption. Studies that meet evidence standards with reservations are strong quasi-experimental studies as well as the randomized trials and regression discontinuity designs having randomization, attrition, or disruption problems. The remaining studies are classified as not meeting evidence screening and thus not providing sufficient evidence for valid causal inference, according to these standards.

Figure 2
Research Designs
(adapted from Campbell & Stanley 1966)

Design Type	Design Name	Visual Representation	In	Ex
Pre- and Non-Experimental Designs	One-Shot Case Study[3]	X O	0	0
	One-Group Pretest-Posttest	O X O	2	0
	Static-Group Comparison	X O O	4	0
True Experimental Designs	Pretest-Posttest Control Group	R O X O R O O	8	0
	Posttest-Only Control Group	R X O R O	8	1
Quasi-Experimental Designs	Nonequivalent Control Group	O X O O O	6	0
	Time Series	O O O O X O O O O	6	0
	Multiple Time-Series	O O O O X O O O O O O O O O O O O	8	0
	Regression Discontinuity	Compares plotted results for groups assigned to treatment at a pre-test cut-off point	6	3

[3] The term "case study" as used by Campbell and Stanley (1966) refers to designs examining one treatment group or program without a comparison group or a pretest. Here it does not imply the use of qualitative as opposed to quantitative methods.

The mapping of studies in terms of outcome variables is less straightforward, particularly in the qualitative studies. The outcomes of a program are defined in multiple ways, and the distinction between process variables and outcome variables is not always clear. In addition, particularly in qualitative studies and exploratory case studies, outcomes variables may emerge as relevant as part of the research process. For the purpose of this review, I chose to identify the variables highlighted in the methods section as an outcome to receive focused attention or a potential result of a mother-tongue program. These variables were listed and analyzed qualitatively using categories that fit the data.

Studies were also classified according to the data collection methods and unit of analysis being targeted, which is typically, but not always, the student. The context of each study was also noted in terms of program type as well as the first and second languages of program participants. Additional information regarding the proportion of articles dealing with evaluation and approaching evidence standards in the various journals was also analyzed.

Results

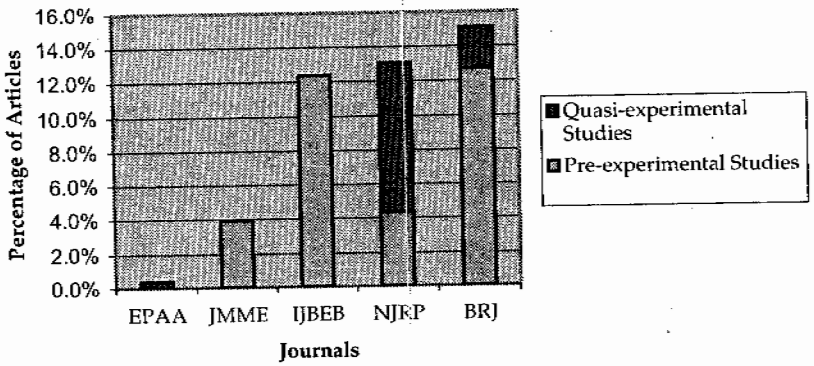
Of the 669 articles published in the selected journals in the past five years, only 41 (6%) were identified as relevant for this review—articles which evaluated or explored the pros or cons of mother-tongue programs. These studies involved evaluation of a variety of programs, many labeled as bilingual education, bilingual immersion, or two-way bilingual immersion. Other programs were labeled as ESL, dual immersion, trilingual education, first-language literacy, transitional bilingual, and developmental bilingual. The programs were mostly at the elementary school level with a few pre-school, middle-school, and high-school programs. Spanish and English were most common as first and second language respectively, although other languages were involved in well over half of the studies. Other first and second language combinations included Chinese and English, Kam and Chinese, Bantu and Portuguese, Rumansch and French, Corsican and French, Austrian and English, Russian and English, Macedonian and Albanian, Hebrew and Arabic, Shipibo and Spanish, Basque and Spanish, and English and Dakota.

Before mapping the articles in terms of research design and outcome measures, I consider the contribution of each selected journal in publishing articles regarding impact evaluation of mother-tongue programs. Figure 3 compares the source journals for this review, showing the percentage of articles published from 2000 to 2004 that evaluate or explore the pros and cons of mother-tongue programs and highlighting the quasi-experimental studies.

The *Bilingual Research Journal* (BRJ) published 126 articles from 2000 to 2004, of which 19 deal with the impact of mother-tongue programs. An

additional 16 articles, not included in this review, focus on the impact of policy changes, specifically the elimination of mother-tongue programs through California's Proposition 227. The National Association for Bilingual Education's *NABE Journal of Research and Practice* (NJRP) began in the winter of 2003. Of the 23 published articles, three relate to mother-tongue program evaluation: two quasi-experimental and one pre-experimental study. The *International Journal of Bilingual Education and Bilingualism* (IJBEB) published 105 articles from 2000-2004, and 13 of these involved mother-tongue program evaluation, all non-experimental case studies and static comparisons. Of the 127 articles in the *Journal of Multilingual and Multicultural Education* (JMME), 5 reported on program evaluations of the non-experimental type, including case studies and a static comparison. The *Education Policy Analysis Archives* (EPAA) contains 272 articles from the past five years, of which only one could be identified as meeting criteria for the review: a quasi-experimental study of bilingual program impact. The *Journal of American Indian Education* (JAIE) was also selected for this review since it has in previous decades published articles dealing with the effectiveness of mother-tongue programs. However, none of the 46 articles published in the past five years directly addresses this topic.

Figure 3
Percentage of Articles in Selected Journals on Impact Evaluation of Mother-Tongue Programs

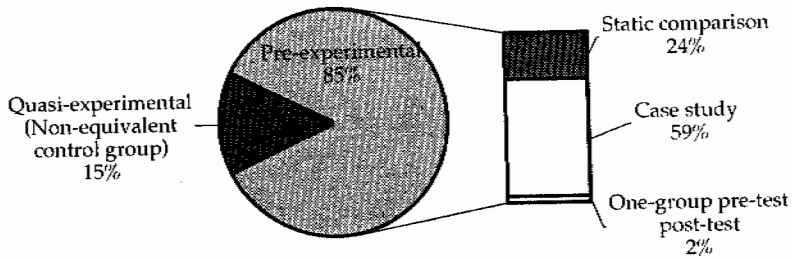


Means of Measurement

Of the 41 studies identified for this review only 6 (15%) could be classified as quasi-experimental, and no true experimental studies were found. These quasi-experimental studies have the potential to "meet evidence standards with reservation" according to the What Works Clearinghouse standards. They were all non-equilavanet control group designs, some using statistical controls to enhance the comparability of the groups. None of the 41 studies would fully meet the What Works Clearinghouse evidence standards.

As Figure 4 demonstrates, the case study is the most common research design. This is understandable since evaluation of one program alone is the simplest design and common particularly for qualitative studies. Still it is surprising that only one of the studies of individual programs included data from a pre-test or pre-program observation. The static comparison, evaluating outcomes of one program in comparison with another using no pretests, is the second most common research design.

Figure 4
Research Designs Used in Evaluating Mother-Tongue Programs

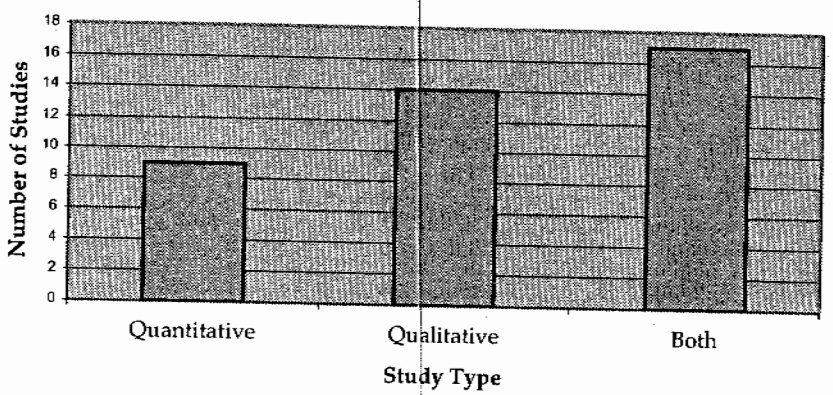


In terms of the type of data collected, the qualitative studies (35%) outnumbered the quantitative studies (23%). However, many studies used both quantitative and qualitative means for addressing the impact of the mother-tongue program, as demonstrated in Figure 5.

The quantitative-qualitative distinction points to the question of how the various outcomes were measured. Almost a fourth of the studies used standardized achievement tests, whether state or other. Several used assessments created for the study itself or borrowed assessments from other researchers. Almost half of the studies used interviews or surveys to gather information from and learn the perceptions of stakeholders. Naturally, observations were also an important means for gathering qualitative data.

As may be expected, the most common unit of analysis was the student. However, 15 of the studies also focused on parents, teachers, or the community. Two studies focused on the impact of the program on the minority language in question. This diversity in units of analysis begin-sto demonstrate the diversity of views regarding relevant outcome measures for evaluating the success of mother-tongue programs.

Figure 5
Studies Focused on Numbers or on Words



Measures of Success

What were the outcome variables used to evaluate the success of the programs in these studies? Appendix A lists the articles identified for this review including outcome variables highlighted in each study. Many of the studies included multiple indicators of program success.

Academic achievement was the most frequent outcome of interest, although this was the primary outcome measure for less than half of the studies. General academic achievement was a focus for about 12% of the studies, literacy for 17%. Near the quantitative end, one study focused specifically on science scores, and towards the qualitative end another study focused on the development of academic discourse.

Almost 20% of the studies included an emphasis on language learning, half of these particularly identifying oral language skills as measures of success. Two additional studies focused on language attitudes and changes in language attitude that may be attributed to the influence of the mother-tongue program.

For those outcome measures focused on academic achievement and language learning, it is important to consider which languages were being evaluated. Often the second language provides the only basis for the outcome variables considered worth measuring. However, 6 of the 41 studies included evaluation of skills in the first language as another important outcome for mother-tongue programs. These measures demonstrate a respect for the first language as more than simply a learning tool.

Teacher, family, and student perceptions regarding the mother-tongue program were important outcome variables in about 22% of the studies. Community factors were also important for about 12% of the studies, including community collaboration brought about by the program, as

well as long-term educational and language maintenance outcomes for the community. Although perceptions and community involvement may be considered part of the process and motivational factors in the programs, positive perceptions and involvement of stakeholders provide an important indication of whether or not a program is successfully meeting needs.

Almost 15% of the studies highlighted identity formation and empowerment as important outcomes of mother-tongue programs. This included changes in the cultural identities of students and teachers, increases in self-confidence, and development of voice by students. Two additional studies focused on the learning environment that was facilitated by the use of the mother tongue in the classroom. Since learning environment, empowerment, and identity formation can hardly be measured by standardized tests, the importance of qualitative research is evident when significant outcome variables such as these are considered.

Discussion

Reactions to current research policies favoring randomized trials and quasi-experimental designs come passionately from all directions. Supporters of these policies allege that most education research is "watery and narrowly descriptive, and does little to inform the public about what types of classroom practices improve student performance" (Glenn 2005: 1). This comment in itself is enough to trigger strong reactions from qualitative researchers not focusing on experimental paradigms. Sound reasons and defenses abound.

"People I work with in evaluation are asking questions that truly cannot be answered through randomized trials." This is a response offered by Sharon Rallis, president of the American Evaluation Association (Glenn 2005: 1). The same issue is highlighted through the various outcome variables compiled in this review. Some of these outcome "measures" simply cannot be measured experimentally or quantitatively. In addition, the exploratory nature of qualitative studies makes them important for identifying relevant variables for assessing program success. These indicators of program success warrant more thorough exploration and analysis. Methods for evaluating these variables, whether qualitatively or quantitatively deserve continued development, building on what has already been done.

The research designs used for classifying studies in this review have been associated with quantitative research. Forcing the categories of quantitative research onto qualitative studies is certainly not ideal. Still the value placed on repeated measures and comparison groups in the high-validity designs provides an important reminder for all researchers. In addition, as mentioned above, the quantitative-qualitative distinction is different from the experimental-non-experimental distinction. A solid-

ly ethnographic study such as Hornberger (1988) may provide a good model as a qualitative version of the multiple time series quasi-experimental design, involving repeated observations of two groups, one with and one without a mother-tongue program. Such studies should be considered worthy of meeting evidence standards just as similar quantitative studies are.

Even with a broadening of evidence standards to include sound qualitative studies, however, few of the articles reviewed could meet current evidence standards in terms of research design. Nor could they contribute to a meta-analysis evaluating the overall value of mother-tongue education. Qualitative researchers may not in fact have this goal in mind. They may be asking other vital questions. Still, qualitative researchers may benefit from an awareness of the threats to the validity of studies and of the options for making studies more acceptable to evaluators and policy-makers.

While random assignment is often unfeasible in ordinary contexts, making true experimental designs difficult, the transformation of a non-experimental design into a quasi-experimental design may be surprisingly simple. Even upgrading a case study to a static group comparison by comparing it with another group or upgrading a one-group pretest posttest design by adding a pretest increases the validity of results. Adding a pretest to that static comparison pulls the study into the quasi-experimental category, controlling for additional threats to internal validity and allowing for more statistical controls. In a school setting multiple unobtrusive measures of achievement over time are not difficult to find. Including multiple pretests and posttests as pre- and post-observations for a single group transforms it into a time-series study. Using multiple pre- and posttests with a comparison group gives the study a multiple time series designs, allowing for interpretation of maturation patterns and the longer-term nature of the impact being measured.

Conclusion

What research designs have been used in evaluating or exploring the pros and cons of mother-tongue programs? The abundance of non-experimental designs and complete lack of true experimental designs used in recent mother-tongue program evaluations is worth noting. Few of the articles reviewed would contribute to a quantitative assessment of the value of mother-tongue education.

What outcome variables have been used as indicators of the success of a mother-tongue program? The outcome measures described above and listed in Appendix A deserve more thorough exploration and analysis. However, the importance of multiple measures, both quantitative and qualitative, of a program's success has been highlighted.

The use of comparison groups and attention to change over time can enhance quantitative and qualitative studies alike. Similarly, evaluations, both quantitative and qualitative, can be enhanced by consideration for the multiple components of program success.

Cynthia Groff is a Ph.D. student in educational linguistics at the University of Pennsylvania's Graduate School of Education. Her research interests include language policy, mother-tongue literacy, multilingual education, and empowerment of linguistic minorities. Her dissertation research focuses on mother-tongue education in India.

E-mail: cgroff@dolphin.upenn.edu

References

Arce, J. (2000). Developing voices: Transformative education in a first-grade two-way Spanish immersion classroom: A participatory study. *Bilingual Research Journal*, 24(3), 225-236.

Alanis, I. (2000). A Texas two-way bilingual program: It's effects on linguistic and academic achievement. *Bilingual Research Journal*, 24(3), 225-248.

Alanis, I., Munter, J., Tinajero, J. (2003). Preventing reading failure for English language learners: Interventions for struggling first-grade L2 students. *NABE Journal of Research and Practice*, 1(1), 92-109.

Amaral, C. (2001). Parents' decisions about bilingual program models. *Bilingual Research Journal*, 25(1&2), 1-23.

Avalos, M. (2003). Effective second-language reading transition: From learner-specific to generic instructional models. *Bilingual Research Journal*, 27(2), 99-205.

Bekerman, Z. & Shhadi, N. (2003). Palestinian-Jewish bilingual education in Israel: Its influence on cultural identities and its impact on intergroup conflict. *Journal of Multilingual and Multicultural Development*, 24(6), 473-484.

Benson, C. (2000). The primary bilingual education experiment in Mozambique, 1993 to 1997. *International Journal of Bilingual Education and Bilingualism*, 3(3), 149-166.

Boruch, R.F. (1975). Contentions about randomized experiments. In R.F. Boruch and H.W. Riecken (Eds.) *Experimental tests of public policy. Boulder, CO: Westview Press.* (Reprinted from *Evaluation studies review annual*, pp. 158-194, by G.V. Glass, Ed., [YEAR] Thousand Oaks, CA: Sage.)

Brohy, C. (2001). Generic and/or specific advantages of bilingualism in a dynamic plurilingual situation: The case of french as official L3 in the school of Samedan (Switzerland). *International Journal of*

Bilingual Education and Bilingualism, 4(1), 38-49.
Campbell, D. & Stanley, J. (1966). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.
Collier, V., & Thomas, W. (2004). The astounding effectiveness of dual language education for all. *NABE Journal of Research and Practice*, 2(1), 1-20.
Cummins, J. (1991). Language development and cognitive learning. In L. Malavé and G. Duquette (Eds.)., *Language, culture and cognition.* Clevedon: Multilingual Matters.
Cummins, J. & Swain, M. (1987). *Bilingualism in education.* New York: Longman.
DeJong, E. (2004). L2 proficiency development in a two-way and a developmental bilingual program. *NABE Journal of Research and Practice 1*(1), 77-108.
DeJong, E. (2004, September 22). After exit: Academic achievement of former English language learners. *Education Policy Analysis Archives, 12*(50). Retrieved December 20, 2004 from http://epaa.asu.edu/epaa/v12n50/v12n50.pdf.
Downes, S. (2001). Sense of Japanese cultural identity within an English partial immersion programme: Should parents worry? *International Journal of Bilingual Education and Bilingualism*, 4(3), 165-180.
Duflo, E. & Kremer, M. (2005). Use of randomization in the evaluation of development effectiveness. In G. Pitman, O.N. Feinstein, and G.K. Ingram (Eds.)., *World Bank Series on evaluation and development: Vol. 7. Evaluating development effectiveness.* (pp. 205-231) New Brunswick, NJ: Transaction Publishers.
Freeman, Y.S., Mercuri, S. & Freeman, D.E.(2001). Keys to success for bilingual students with limited formal schooling. *Bilingual Research Journal, 25*(1&2), 1-11.
Geary, D. & Pan, Y. (2003). A bilingual education pilot project among the Kam people in Guizhou Province, China. *Journal of Multilingual and Multicultural Development, 24*(4), 274-289.
Ghadessy, M. (2002). Attitude change in bilingual education: The case of Brunei Darussalam. *International Journal of Bilingual Education and Bilingualism*, 5(2), 113-128.
Glenn, D. (2005, February 24). New federal policy favoring randomized trials in education research takes effect today. *The Chronicle of Higher Education.* Retrieved March 8, 2005 from http://chronicle.com/prm/daily/2005/02/2005022401n.htm
Griessler, M. (2001). The effects of third language learning on second language proficiency: An Austrian example. *International Journal of Bilingual Education and Bilingualism*, 4(1), 50-60.
Holmes, D., & Duron, S. (2000). *LEP students and high-stakes assessment.* NCELA Report. Retrieved December 20, 2004 from http://www.ncela.gwu.edu/pubs/reports/highstakes/index.htm.
Hornberger, N. (1988). *Bilingual education and language maintenance: A southern Peruvian Quechua case.* Dordrecht, Holland: Foris.
Hornberger, N. (1991). Extending enrichment bilingual education: Revising typologies and redirecting policy. In O. García (Ed.) *Bilingual education Focusschrift in honor of Joshua A. Fishman* (Vol. 1: 215-234). Philadelphia: John Benjamins.
Housen, A. (2002). Processes and outcomes in the European schools model of multilingual education. *Bilingual Research Journal, 26*(1), 1-20.
Hovens, M. (2002). Bilingual education in West Africa: Does it work? *International Journal of Bilingual Education and Bilingualism*, 5(5), 249-266.
Jaffe, A. (2003). Talk around text: Literacy practices, cultural identity and authority in a Corsican bilingual classroom. *International Journal of Bilingual Education and Bilingualism*, 6(3), 202-220.
Johnson, R.J. (2000). Case studies of expectation climate at two bilingual education schools. Bilingual Research Journal, 24(3), 225-240.
Johnston (2002). Case studies of expectation climate at two bilingual education schools. *Multilingual and Multicultural Development*, 24(3), 195-213.
Jong, E. (2002). Effective bilingual education: From theory to academic achievement in a two-way bilingual program. *Bilingual Research Journal, 26*(1), 1-20.
Kemppainen, K., Ferrin, S.E., Ward, C.J., & Hite, J.M. (2004). "One should not forget one's mother tongue": Russian-speaking Parents' choice of language of instruction in Estonia. *Bilingual Research Journal, 28*(2), 207-229.
Kirk Senesac, B.V. (2002). Two-way bilingual immersion: A portrait of quality schooling. *Bilingual Research Journal, 26*(1),1-17
Lao, C. (2004). Parents' attitudes toward Chinese-English bilingual education and Chinese-language use. Bilingual Research Journal, 28(1), 99-121.
Lasagabaster, D. (2001). Bilingualism, immersion programmes and language learning in the Basque Country. *Journal of Multilingual and Multicultural Development*, 22(5), 401-425.
Lemberger, N. (2002). Russian bilingual science learning: Perspectives from secondary students. *International Journal of Bilingual Education and Bilingualism*, 5(1), 58-71.
Leow, C., & Boruch, R.F. (2001). *Locating randomized experiments on math and science education: A hand search and machine based searches of the American Education Research Journal.* Report of the Campbell

Collaboration Secretariat. Philadelphia: University of Pennsylvania, Graduate School of Education.

López, M. & Tashakkori, A. (2004). Effects of a two-way bilingual program on the literacy development of students in kindergarten and first grade. *Bilingual Research Journal, 28*(1), 19-34.

López-Bonilla, G. (2002). Los programas de inmersión Bilingüe y la adquisición del discurso académico. *Bilingual Research Journal, 26*(3), 525-536.

Manyak, P. (2002). "Welcome to Salon 110": The consequences of hybrid literacy practices in a primary-grade English immersion class. *Bilingual Research Journal, 26*(2), 213-234.

McDonald, S., Lefebvre, C., Antes, G., Galandi, D., Gøtzsche, P., Hammarquist, C., Haugh, M., Jensen, K.L., Kleijnen, J., Loep, M., Pistotti, V., Rüther, A. (2002). The contribution of handsearching European general health care journals to the Cohcrane Controlled Trials Register. *Evaluation & the Health Professions, 25*(1): 65-75.

Mosteller, F. & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education research.* Washington, D.C.: The Brookings Institute.

Mulhem, M. (2002). Two kindergartners' constructions of literacy learning in Spanish: A challenge to superficial balanced literacy instruction. *International Journal of Bilingual Education and Bilingualism, 5*(1), 20-39.

Ordonez (2004). EFL and native Spanish in elite bilingual schools in Colombia: A first look at bilingual adolescent frog stories. *International Journal of Bilingual Education and Bilingualism, 7*(5), 449-474.

Rossi, P., Lipsey, M., & Freeman, H. (2004). *Evaluation: A systematic approach* (3rd ed.). Thousand Oaks, CA: Sage.

Shannon, S.M. & Milian, M. (2002). Parents choose dual language programs in Colorado: A survey. *Bilingual Research Journal, 26*(3), 681-696.

Shavelson, R. and Towne, L. (Eds) (2002). *Scientific research in education.* Washington, DC: National Academy Press, 2002.

Sheffer, C.S. (2003). Parents' lack of understanding of their children's bilingual education program. *Bilingual Research Journal, 27*(2), 333-341.

Smith, P.H., Arnot-Hopffer, E., Murphy, E., Valle Davis, A., Gonzalez, N. Poveda, A. (2002). Raise a child, not a test score: Perspectives on bilingual education at Gavis Bilingual Magnet School. *Bilingual Research Journal, 26*(1), 1-19.

Spezzini (2004). English immersion in Paraguay: Individual and sociocultural dimensions of language learning and use. *International Journal of Bilingual Education and Bilingualism, 7*(5), 412-431.

Tankersley, D. (2001). Bombs or bilingual programmes?: Dual-language immersion, transformative education and community building in Macedonia. *International Journal of Bilingual Education and Bilingualism, 4*(2), 107-124.

Thompson, M.S., DiCerbo, K.E., Mahoney, K. and MacSwan, J. (2002, January 25). Exito en California? A validity critique of language program evaluations and analysis of English learner test scores. *Education Policy Analysis Archives, 10*(7). Retrieved December 20, 2004 from http://epaa.asu.edu/epaa/v10n7/.

Torres-Guzmán, M., Abbate, J., Estela, M., & Minaya-Rowe, L. (2002). Defining and documenting success for bilingual learners: A collective case study. *Bilingual Research Journal, 26*(1), 1-19.

Turner, H., Boruch, R., Petrosino, A., Lavenberg, J., DeMoya, D. & Rothstein, H. (2003). Populating an international web-based randomized trials register in the social, behavioral, criminological, and education sciences. *Annals of the American Academy of Political and Social Science, 589*, 203-223.

Wilde, E. & Hollister, R. (2002). *How close is close enough: Testing non-experimental estimates of impact against experimental estimates of impact with education test scores as outcomes.* Institute for Research on Poverty, University of Wisconsin, Discussion paper No. 1242-02. Retrieved March 8, 2005 from http://www.irp.wisc.edu/publications/dps/pdfs/dp124202.pdf

Yamauchi, L., Ceppi, A. & Lau-Smith, J. (2000). Teaching in a Hawaiian context: Educator perspectives on the Hawaiian language immersion program. *Bilingual Research Journal, 24*(4), 385-403.

Yip, D.Y., Tsang, W. K., Cheung, S.P. (2003). Evaluation of the effects of medium of instruction on the science learning of Hong Kong secondary students: Performance on the science achievement test. *Bilingual Research Journal, 27*(2), 295-331.

Zhou, M. (2001). The politics of bilingual education and educational levels in ethnic minority communities in China. *International Journal of Bilingual Education and Bilingualism, 4*(2), 125-149.

Appendix A
Studies Evaluating the Impact of Mother-Tongue Programs

	Author	Year	Journal	Outcome Variables
1	Acre	2000	BRJ	Development of student's voice in dialogue; Community of learners.
2	Alanis	2000	BRJ	First and second language
3	Alanis et al.	2003	NJRP	Reading
4	Amaral	2001	BRJ	Parent perceptions of student achievement in different programs
5	Avalos	2003	BRJ	Comprehension: Written Recalls
6	Bekerman & Shhadi	2003	JMME	Cultural identities, student's conceptions; Inter-group conflict resolution
7	Benson	2000	IJEEB	Classroom participation, self-confidence, bilingualism, biliteracy
8	Brohy	2001	IJEEB	Language attitudes; Language competence
9	Collier & Thomas	2004	NJRP	Student achievement; Teacher, administrator & parent perspectives
10	DeJong	2004a	EPAA	Academic achievement
11	DeJong	2004b	NJRP	Second-language proficiency
12	Downes	2001	IJEEB	Cultural identity, attitude towards culture
13	Freeman et al.	2004	BRJ	Learning community, learning opportunities, engagement of students
14	Geary & Pan	2003	JMME	First-language reading and writing; Class participation
15	Ghadessy	2002	IJEEB	Student attitudes; English reading & comprehension
16	Griessler	2001	IJEEB	Oral picture book narrations
17	Housen	2002	BRJ	Academic achievement
18	Hovens	2002	IJEEB	Test scores; Teacher-pupil interaction
19	Jaffe	2003	IJEEB	Identities for minority language and learners
20	Johnson	2000	BRJ	Expectation climate
21	Johnston	2002	JMME	First-language mastery
22	Jong	2002	BRJ	Academic achievement: Eng. and Span. comprehensive assessments
23	Kemppainen et al.	2004	BRJ	Parent opinions
24	Kirk Senesac	2002	BRJ	Achievement; Parent & community collaboration
25	Lao	2004	BRJ	Parent opinions
26	Lasagabaster	2001	JMME	Language shift: Linguistic and non-linguistic results
27	Lemberger	2002	IJEEB	Student perceptions
28	Lopez & Tashakkori	2004	BRJ	Literacy development
29	Lopez-Bonilla	2002	BRJ	Academic discourse
30	Manyak	2002	BRJ	Use of linguistic repertoire as a resource, developmental biliteracy
31	Mulhem	2002	IJEEB	Understanding about written language
32	Ordonez	2004	IJEEB	Oral narrative proficiency
33	Shannon & Milian	2002	BRJ	Parent opinions
34	Sheffer	2003	BRJ	Parents' understanding and expectations for English instruction
35	Smith et al.	2000	BRJ	Teacher, family, student perspectives
36	Spezzini	2004	IJEEB	English comprehensibility; Student perspectives
37	Tacelosky	2001	JMME	Language use
38	Tankersley	2001	IJEEB	Community-building, conflict resolution
39	Yamauchi et al.	2000	BRJ	Teachers' transformation of identities as educators and as Hawaiians
40	Yip et al.	2003	BRJ	Science achievement test scores
41	Zhou	2001	IJEEB	Community education level