Measuring the Efficacy of *Big Math for Little Kids*: A Look at Fidelity of Implementation[1]

Barbrina Ertle and Herbert P. Ginsburg

Teachers College, Columbia University

Ashley E. Lewis

EDC Center for Children and Technology

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 9, 2006

[1] This work was funded by the Institute for Education Sciences, US Department of Education under award R305K040001. The opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Institute for Education Sciences. The authors would like to acknowledge the contributions of their colleagues on the project – Ellen Mandinach, Melissa Morgenlander, Leslie Manlapig, and Maria Cordero.

Measuring the Efficacy of *Big Math for Little Kids*:

A Look at Fidelity of Implementation

Since the inception of the No Child Left Behind Act (NCLB), the need and desire for research-based curricula has increased substantially. Not only do these curricula need to reflect previous research findings from developmental and cognitive science, but each curriculum itself needs to undergo rigorous scientific evaluation to show that it "works." Shavelson and Towne (2002) and the What Works Clearinghouse (2004) describe and advocate for these rigorous research standards in the hopes that these methodologies will authoritatively indicate "what works." Often, this vetting process involves randomized controlled trials (RCTs) to show that the curriculum produces strong, positive learning outcomes in diverse groups of students (Brass, Nunez-Neto, & Williams, 2006).

RTCs involve random assignment of subjects, exposing the intervention only in the treatment condition, and averaging the outcome measure of each group to determining if differences between groups are due to chance (Brass et. al., 2006). However, even this experimental methodology many not be enough (Mowbray, Holter, Teague, & Bybee, 2003). In order for a research study to examine the impact of a curriculum, it must ensure that the curriculum is implemented as the designers intended (Mowbray et. al., 2003). Thus, researchers need a tool that sensitively measures the degree to which teachers adhere to the intervention, in this case a curriculum. This paper discusses the concept of implementation fidelity, details our process of developing and using a fidelity measures, and explores way in which our experience can generalizing beyond our study and inform other researchers.

Mowbray and colleagues (2003) describes fidelity as "the extent to which delivery of an intervention adheres to the protocol or program model originally developed (p. 315)." More

specifically, Borrelli, Sepinwall, Ernst, Bellg, Czajkowski, Breger, DeFrancesco, Levesque, Sharp, and Ogedegbe (2005) state that:

> Treatment fidelity refers to the methodological strategies used to monitor and enhance the reliability and validity of behavioral interventions. The overall goal of enhancing treatment fidelity is to increase scientific confidence that changes in the dependent variable are attributable to the independent variable (p. 852).

In essence, high fidelity teachers remain loyal to the necessary components of the curriculum, which is expected to lead to increases in the dependent variable, which in educational contexts is usually learning. Tracking fidelity actually helps ensure that the research qualifies as an efficacy trial, which seeks to determine if the independent variable actually affects the dependent variable "under ideal conditions (Shadish, Cook, & Campbell, 2002, p. 507)." Ideal conditions are synonymous with high fidelity; without measuring implementation fidelity, efficacy trials are simply assuming that they are studying the treatment under ideal conditions.

The degree to which the curriculum is implemented with fidelity is also important for accurate interpretation of outcome data (Mowbray et. al., 2003). It does this in three specific ways: it (1) explains the study's findings, (2) helps revise interventions, and (3) increases the statistical power and effect size because it reduces unwanted variability (Borrelli et. al. 2005). It can also be used as a moderating variable in statistical analyses (Mowbray et. al., 2003). If the data analysis shows significant differences between the experimental and control groups, high fidelity helps to explain these findings by supporting the idea that the experimental group actually experienced the treatment, while low fidelity might suggest that something other than the treatment led to positive outcomes. However, if the results show no differences there are two possible explanations. It is possible that the treatment is actually not effective, a conclusion supported by high fidelity. Low fidelity suggests that the treatment may be effective with proper levels of implementation, but that non-significant results should not be used as evidence that the

implementation does not work. If properly used, fidelity information provides insight and decrease the likelihood of both Type I and Type II errors (Borrelli et. al., 2005).

Fidelity also informs the way research and evaluation studies are assessed for quality by other researchers and policymakers (Brass et. al., 2006). Studies that maintain high fidelity are more trustworthy because fidelity insures both internal and external fidelity (Borrelli et. al., 2005). High fidelity increases internal validity because it means that the researchers are actually studying what they intend to study. In turn, the external validity increases because given a high level of fidelity, the observed outcomes are likely to replicate in other schools as well. Generalizability, or the ability to assume that the same causal relationship exists in settings outside the context of the experiment itself (Shadish et. al., 2002), is a major justification for using randomized control trials in the first place. External validity and generalizability are often synonymous terms (Borrelli et. al., 2005). Measuring fidelity of implementation helps to ensure that experimental designs provide results usable in other educational contexts, an underlying purpose of the What Works Clearinghouse (2004).

In many ways, fidelity of implementation is similar to construct validity, a third type of validity. Construct validity links the definition of the construct with the construct itself to ensure that the definition actually captures the construct it claims to (Shadish et. al., 2002). Within the program evaluation, construct validity often refers to the way the program is intended to operate, as well as the causal chain connecting the program to expected outcomes (Brass et. al., 2006). In our case, construct validity is established by showing that the definition the program, as illustrated by the fidelity measures, is accurately capturing the curriculum's construct.

Assessing the level of fidelity is important for methodological reasons, but also has important implications for practitioners. Fidelity of implementation has potential relevance for

school administrators and curriculum specialists. However, research results are not as easily usable to practitioners as researchers might expect. Detailed and data-heavy research reports presented information in formats that are unsuitable for administrators, practitioners and policymakers, who often start by asking questions like “Does it work?” and really want to know “Will it work for *my teachers or district?”* (M. Honey, personal communication, November 30, 2005). These stakeholders want to know whether the curriculum leads to positive learning outcomes, which randomized control field trials can inform. More importantly, they want to know whether the curriculum will work for their school. Fidelity can help answer these questions by indicating what it takes to successfully implement the curriculum; however, this requires researchers to break down fidelity information and organize it into a format usable by practitioners. For example, if a curriculum requires students to use many materials, potential curriculum adopters need to take the cost of these materials into account.

Fidelity criteria could also directly impact teachers by informing the training process and providing the foundation for teacher assessment, as some researchers recommend (Borrelli et. al., 2005). For example, fidelity measures focus on key components of the curriculum, components which could also be used to train and guide teachers. Teachers who know what aspects of a curriculum are critical and why they are critical are more likely to implement with fidelity and achieving the learning goals set out by the curriculum. This is especially important for the sustainability of the curriculum, particularly for schools that were once under the watchful eye of researchers. In the future, fidelity criteria could be translated into a teachers-friendly format that might be used for self- or peer-assessment.

The “Big Math for Little Kids” (BMLK; Ginsburg, Greenes, & Balfanz, 2003) curriculum is designed for preK and kindergarten children and is not yet bound by the tenets of

NCLB; however, the need to establish the program's efficacy is no less important. In order to properly determine this, we needed to develop a valid and reliable fidelity measure and employ it at several points throughout the curriculum. In order to create a fidelity measure it is important to fully understand the curriculums goals, activities, and structure. Necessary components of the curriculum are then extrapolated, defined, and quantified.

Admittedly, developing these measures, validating them, and using them is time intensive and costly; however, the cost recommending a program that is not really effective or only effective if used in a particular way is expensive too (Borrelli et. al., 2005). If the program works, teachers that demonstrate these key components with greater loyalty are then expected to produce better outcomes than teachers who implement the curriculum with less accuracy. Ultimately, it is not enough that the curriculum is used, but rather that it is used accurately and appropriately as the curriculum developers originally intended.

"Big Math For Little Kids": Math for Young Children

Big Math for Little Kids is a comprehensive mathematics curriculum developed for preK and kindergarten children. At each level, the curriculum offers a structured sequence of activities designed to promote challenging mathematical learning and related verbal expression in six major content strands: number, shape, measurement, operations on numbers, patterns and logic, and space (Ginsburg et. al., 2003). For each of these strands, *BMLK* also offers storybooks that illustrate key math concepts and provide the associated mathematical language.

BMLK was developed with NSF support based on research findings indicating that young children are ready to learn math, capable of learning math, and enjoy learning math. Founded on

these principles, *BMLK* aims to provide young children with an enjoyable, meaningful, and rich mathematical experience that lays the groundwork for academic success.

Measurement of Curriculum Efficacy

Pilot research and field-tests indicate that the curriculum is effective for lower-SES children, and indeed for children from all social backgrounds, helping them to achieve high levels of mathematics learning and to improve their language skills. The study described here involves evaluation of the curriculum in a more rigorous fashion using a random assignment design for approximately 640 children in preK classrooms and 320 children in kindergarten classrooms.

Our central hypothesis for this study is that in preK and kindergarten, *BMLK* promotes more extensive mathematics learning than does a control group experience. This is being tested using a measure of student outcomes, which is of obvious importance for evaluation of a mathematics curriculum. But although measuring student achievement can provide some indication of the effect of the *BMLK* curriculum, it does so without provision for differences in the ways teachers implement the curriculum. Therefore, we also want to determine the degree to which student outcomes are affected by fidelity of implementation. Our hypothesis is that children's performance will be positively related to the fidelity with which teachers implement the program. This paper addresses this second aspect of the study – the measurement of implementation fidelity.

Given that a measure of fidelity must be specific to the program being implemented, we had two initial tasks: one, to develop a measure of fidelity for the *BMLK* curriculum; and two, to identify measures of fidelity (if they existed) for the program(s) utilized in control group

classrooms. A survey of New York City's Administration for Children's Services (ACS) childcare centers, from which the study groups were selected, found that the majority of centers utilized the *Creative Curriculum* (*CC*; Dodge, Colker, Heroman, 2002), with additional centers utilizing *High Scope* (*HS*) (Hohmann & Wiekart, 2002) or a "developmentally appropriate" approach to teaching preschool (in which no particular curriculum is specifically followed).

To begin developing a measure of fidelity for *BMLK*, we looked at the 'fidelity' measures for the possible curricula that would likely be used by the control group teachers – *Creative Curriculum*'s Implementation Checklist (IC), and *High Scope*'s Preschool Program Quality Assessment (PQA)[2]. Both of these curricula are based on philosophies of engaging in teaching based on children's play (Dodge et. al., 2002; Hohmann, 2002). Both of these curricula include goals for mathematics learning. In examining their 'fidelity' measures, though, we found that they really did not measure the teaching that the curricula were promoting, especially in relation to mathematics, except with very broad items that would not allow 'unpacking' of the characteristics of implementation with which teachers complied or did not comply. Such 'unpacking' would be essential to allow us to determine what aspects of implementation are most positively related to improved student outcomes – a necessary understanding to promote teacher improvement and effective professional development.

Looking for Something Different: Fidelity and its Development Process

Given this finding, we decided to take a different approach to developing our fidelity measure. We began our development process by conducting two parallel activities. First, we reviewed field notes from *BMLK* observations, to identify characteristics that we predicted

[2] With no specific curricula being identified with those centers using a "developmentally appropriate" approach, no fidelity measures were examined in relation to these centers.

would impact the quality of children's learning. Second, we began reviewing available measures developed to examine elements of teaching quality, as we predict that it is the elements of fidelity related to quality teaching that would be the most predictive of student outcomes.

Our parallel reviews resulted in a very long list of elements we felt would impact the children's math learning. We were then left with the task of trying to pare down those elements into a practical measure, trying to walk a delicate balance between practicality and validity. The balance required us first to accept the impossibility of being able to examine and measure all aspects of teaching that impact quality. Second, we had to accept that we had proposed to measure fidelity of implementation, which is not necessarily a measure of teaching quality. Fidelity means 'accuracy, loyalty, or faithfulness', and would indicate the extent to which teachers 'followed' a curriculum as it is written. It provides no insight into the quality of that curriculum, or of the quality of learning that would result from its use. Indeed, there is a sense in which strict fidelity is the enemy of quality because the latter requires creative adaptation to the idiosyncratic needs of students and the vagaries of the current situation. And quality bears no affiliation to any specific curriculum, but rather, to the specific teaching that is done in the classroom.

Another, though separate, consideration was a goal we had to measure the 'spirit' of the implementation versus a strict, 'by the letter' implementation. Not only can curriculum developers not foresee and plan to meet the exact needs of every classroom, but to attempt to do so would be demeaning to the teaching profession. Teachers need to be able to think for themselves, to flexibly adapt to the changing needs of their classroom and their children. They need to be able to improvise when specified materials are not available. They need to utilize the space they have available, and to navigate elements over which they have no control. Given the

uniqueness of every classroom, the classroom teacher has to determine what may work best in her classroom. Also, talented teachers can develop adaptations that exceed the vision of curriculum developers, or find improvements on various aspects of any given activity. Therefore, we wanted to be able to develop a measure that would capture an element of 'flexibility' and 'adaptation', to allow for good teaching to evolve. The measure we finally arrived at included six basic items, which could be defined precisely for any *BMLK* activity.

Content Coverage. The first item addresses content coverage – its accuracy, coherence, and alignment with the *BMLK* activity goals. It allows for flexibility in achieving the goals, so far as the goals are achieved in an accurate and coherent manner. For example, in an activity called Bag It, the teacher provides children with empty Ziploc bags labeled 0 to 3, one bag for each number. The teacher then provides a variety of counting objects and children count out the appropriate number of objects and put inside each Ziploc bag, then place each bag into a bin similarly labeled. The content of the activity includes the numbers zero through three, connecting number words with numerals and numbers of objects, and the idea that the last number counted tells how many in a group. Therefore, the teacher would be rated on how well this specific content was addressed, regardless of the manner in which it was addressed.

Teacher's Directions. The second item of the measure addresses the directions given by the teacher – their clarity, appropriateness, and accuracy in alignment with *BMLK* activity guidelines. This item allows some flexibility in the tasks performed, as long as they do not detract from the carefully designed activity. For example, in the Bag It activity, teachers would be rated on how well they gave directions for accomplishing the task, the extent to which they generally adhered to the guidelines, and the extent to which any adaptations contributed to or detracted from the intended tasks.

Children's Engagement. The third item addresses children's engagement – although not directly a measure of fidelity, this item is useful as an indicator of qualitative aspects of the teachers' implementation. Also, children's engagement should undoubtedly be related to student outcomes, as students need to attend to an activity in order to learn. It must be noted, though, that engagement does not necessarily indicate students doing the task as directed, or even doing the task at all, but giving it their attention.

Materials. The fourth item of the measure addresses materials – the appropriateness, sufficient numbers, and alignment with *BMLK* activity guidelines. This allows for flexibility in choices of materials, so far as the selected materials afford the same mathematical learning. For example, in the Bag It activity, teachers would be rated on whether the materials they selected afforded the mathematical goals, and were appropriately used to do so. Namely, teachers were rated on whether they provided: a transparent bag labeled with a number that would allow children to connect the written number with the number of objects in the bag, a variety of appropriate objects for counting and placing in the bags, and bins labeled with each number that would further reinforce the connection with the written number.

Vocabulary. The fifth item of the measure addresses vocabulary – given the curriculum's focus on developing language, appropriate mathematical vocabulary is provided. For example, the vocabulary and mathematical language provided for Bag It includes: "count from 0 to 3", "count to tell how many", and "0, 1, 2, 3" in reference to the written numeral, the number word, and the number of objects in a set. Teachers would be rated on whether they used the specified vocabulary, or elicited the vocabulary from the children.

Global Rating. The sixth item, a more ambiguous 'global rating', was our concession to the elimination of other aspects of quality not included in the measure. This rating subsumes all

previous items, but allows the rater to consider other elements not directly addressed by the measure that could have a significant impact on the overall efficacy of the activity.

Studying Fidelity: The Challenges and Opportunities

Once this measure was developed, the task of implementing it began. Piloting was done to carefully define the different items for the activities selected for observation, and to train observers to use the measure. Given that the process of data collection (both fidelity and student outcomes) are still incomplete for the first cycle of this study, it is too early to delve into findings related to the use of the measure.

However, the process of development and implementation has brought to light numerous challenges associated with developing such a measure. These challenges may be just as important to share with the community as are the findings we will eventually report. In this section we will describe some of the challenges we faced, and how we have chosen to deal with some of them. These seven challenges are presented in no specific order of importance.

Challenge 1: Incomparability

Our overall goal for this study is to measure the efficacy of *BMLK* as a means of promoting understanding of mathematics among young children. Given that our study involves random assignment of children to two different curricula, the element of relating fidelity to student outcomes is a bit complicated. Fidelity of implementation to student learning can only be examined within each program; it is not possible to compare implementation effectiveness between *BMLK* and the control programs because they are so different from each other.

Even accepting the lack of comparability, though, the design of a study that includes fidelity measurement for two different groups involves more subtle challenges. For this study, we proposed observing four *BMLK* activities and four control activities – for fidelity measurements. Selecting four *BMLK* activities was done carefully and rationally, in an attempt to sample various content and various types of activities. However, selection of parallel control observations proved more difficult.

Neither *Creative Curriculum* (CC) nor *High Scope* (HS) are highly structured curricula (Dodge, Colker, Heroman, 2002; NAEYC & NCTM, 2002). Both recommend that teachers use spontaneous play to promote learning. The philosophy underlying both curricula is that play is the most effective strategy to help young children grow intellectually, socially, and emotionally. Given this approach, although the curriculum guides provide suggested activities, if teachers were to maintain true fidelity to the curricula, there would be no way of knowing whether a scheduled observation would involve any mathematical teaching. This presented a challenge in being able to observe and measure mathematical fidelity while maintaining balanced attention to both treatment and control groups.

Also, with the random selection and assignment of eligible centers into the treatment and control groups, the control group includes centers using *CC*, *HS*, and the "developmentally appropriate" approach. This presents the challenge of rating each center with the fidelity measure appropriate to their choice of curriculum. And added to this challenge is the fact that no 'fidelity measure' exists for the unspecific "developmentally approach".

Challenge 2: Finding Comparability

Given that fidelity cannot directly be compared between groups, or even within the control group, we had to look beyond fidelity to find ways to directly compare the two groups of teachers. We turned our attention to developing a general measure of mathematical teaching quality, which could be used for any math activity or lesson. Although this requires us to have control groups 'teach' a math activity – which doesn't necessarily comply with fidelity to their curricula – a survey of the teachers in the group showed that they all 'claim' to teach mathematics to their children through planned activities.

Other measures – examining specific teachers' knowledge regarding content and student learning, and their assessment of student understanding – have also been developed to target specific mathematical concepts included in *BMLK*, *CC*, and *HS*.

Challenge 3: Quality vs. Fidelity

The development of a quality measure then brought us full circle, back to one of our original struggles – separating quality from fidelity. Although our intention of developing the first measure was to examine fidelity, our ideology and aspiration is for quality teaching, which is the same ideology upon which the curriculum was developed. So although quality and fidelity are not the same, their measurement becomes more equivalent the closer aligned the curriculum is to indicators of quality. If our primary hypothesis is true, and if *BMLK* does provide a quality mathematics curriculum, then fidelity to the curriculum should be synonymous to quality mathematics teaching.

Thus, we are mentally focused on quality, and often feel ourselves being distracted by elements of quality unrelated to fidelity. We are finding that we must frequently take a 'step back' to stay focused on the purpose of this study, reminding ourselves of the difference between

fidelity and quality. The introduction of the quality measure makes that challenge even more difficult, as we try to navigate between measuring elements of fidelity vs. quality.

Challenge 4: Defining the Measure

Another challenge comes in defining the fidelity measure for the individual activities for which it is being used. The better defined the ratings are, the easier it is to use the measures with confidence and reliability. But the better defined they are, the more it promotes examining elements of the teaching in isolation, and moving the focus more towards 'by-the-letter' implementation instead of the intended 'spirit' of implementation. It removes the ability to allow for flexibility and adaptation, both of which may improve the quality of the activity instead of worsen it (as strict ratings would imply). Wanting to achieve both flexibility and reliability has also required a balancing act.

Challenge 5: Measure Subjectivity

The issue regarding the subjective nature of defining quality must also be mentioned, if but briefly. Given our belief that the *BMLK* curriculum is the quality curriculum that we hypothesize it to be, quality and fidelity should be synonymous. We acknowledge the elements of quality/fidelity that we have chosen to examine may not be in agreement with those that others would have us examine.

Challenge 6: Qualified Observers

Another challenge results from the flexibility we have built into our measure. It results in the need to have qualified observers to conduct the observations, as ratings are subjective. The

qualifications we speak of, though, are not qualifications for which we can necessarily train. Observers need to be not only familiar with the activities, but have a strong mathematical understanding from which to be able to recognize appropriate adaptations and achievement of the mathematical goals.

As part of this study, we have also been developing and presenting professional development workshops for the teachers in the treatment group – other curricula offer professional development for their programs also. These workshops all include a strong focus on the mathematical content of the different strands. As a result, our group has engaged in long and deep conversations about the mathematical content. Although the purpose of so doing was to develop the workshops, they also served to provide mathematical training for conducting observations.

Challenge 7: Considering Teacher Feelings

Another challenge regarding both the design and the implementation of these measures is the general apprehension and anxiety among teachers about being 'evaluated'. We have spent many hours discussing how to word our questions, introduce our expectations to participating teachers, and implementing the measures so as to make the teachers more comfortable, assure them that we can only succeed in improving early mathematics education with their help, and treat them as the professionals that they are.

Implications for Research and Practice

Emphasis on measuring implementation fidelity as part of RCTs has increased, rightfully so. High fidelity increases the internal, external, and construct validity and can aid in the proper

interpretation of statistical results. Just as medical researchers need to confirm that patients actually experience the treatment level under investigation, educational researchers must confirm that teachers actually provide the curriculum to child in the way curriculum developers think they should. Without this information, researchers cannot be sure that the causal links that RCTs suggest are due to the construct under investigation. Although fidelity is costly to measure, suggesting curricula that do not lead to the desired results in other settings wastes educators time and effort, taxpayers money, and robs children of the best education we can provide.

In this study, we are investigating whether the *BMLK* curriculum leads to better mathematical understanding in young children, preK and kindergarten in particular. Although data analysis is still premature, we are planning to include fidelity results as a moderating variable in our Hierarchical Linear Model (HLM). We want to corroborate both that the curriculum was implemented effectively and determine if student outcomes differ when fidelity is lower than expected. Ultimately, we want to know the degree of fidelity needed to obtain in order to improve students' mathematical understanding.

The first implication of our study is that despite this challenging and time-consuming nature of the development process, fidelity of clear curricula is possible. The development process itself requires deep understanding of the curriculum's key components and a clear idea of how these components differ in real classrooms. Frequent reflection and revision of the measure is needed to make it clear, define the components, and provide indicators of each level of the measure's scale.

The second implication is that it is not reasonable or advisable to measure the fidelity of a curriculum that does not have clear key components. While it is possible to create a measure of general components of non-specific curricula, it is unclear what the results of this measure

actually capture. Researchers should carefully consider whether their intervention is sufficiently specific before attempting to measure fidelity.

The third implication is that comparing curricula on fidelity is difficult, particularly with a non-specific curriculum, and thus fidelity may not be an appropriate way to compare groups. If researchers determine that it is appropriate to compare curricula on their respective measures of fidelity, these results need to be placed on an equivalent scale. In our case, there was not a comparable fidelity measure for the control group. Instead, we created a measure of quality teaching based on the National Association for the Education of Young Children (NAEYC) and the National Council for Teachers of Mathematics (NCTM) standards (2002). Although quality is clearly different from fidelity, it was necessary to have some indicator that would allow us to compare teachers to one another in a meaningful way in our statistical model (HLM). Struggles to compare groups is likely to occur on other research projects as well and although this remains a challenge with no clear answers, the approach we took is one way to tackle the dilemma. In the end, we used the fidelity measure in our treatment classes only and the quality measure in both the treatment and control classes.

The fourth implication is that fidelity measures also need to be grounded in practicality and contextually appropriate. While empirically sound, it may not be feasible to measure fidelity as often as researchers may wish. The number of fidelity ratings depends on several factors, including time, number of raters and their expertise, access to classrooms, and the length of the measure itself. In addition, the fidelity measure needs to be flexible enough to allow teachers to adapt and differentiate the curriculum to meet the specific needs of their students. These deviations from the “letter” of the curriculum conceivable threaten the fidelity rating, yet these adaptations are often necessary and desirable in real-world settings.

Fidelity's role in RCTs is increasing for a number of good reasons, as it provides a particular type of information. However, many things should be considered when researchers are planning to assess fidelity. The overall goal of research is to determine what interventions lead to the best educational outcomes. Holistically, our greater goal is to improve the early childhood mathematics teaching and learning and provide the momentum to help children succeed in the years to their subsequent academic career.

References

Borrelli, B., Sepinwall, D., Ernst, D., Bellg, A. J., Szajkowski, S., Breger, R., DeFrancesco, C., Levesque, C., Sharp, D. L., & Ogedegbe, G. (2005). A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting and Clinical Psychology, 73 (5),* 852-860.

Brass, C. T., Nunez-Neto, B., & Williams, E. D. (2006). *Congress and program evaluation: An overview of randomized controlled trials (RCTs) and related issues*. Washington, DC: Congressional Research Service, Library of Congress. Retrieved March 14, 2006, from http://opencrs.cdt.org/document/RL33301

Dodge, D. T., Colker, L. J., & Heroman, C. (2002). *The creative curriculum for preschool* (4th ed.). Washington, DC: Teaching strategies.

Ginsburg, H.P., Greenes, C., & Balfanz, R. (2003). *Big Math for Little Kids.* Parsippany, NJ: Pearson Education, Inc.

Hohmann, M. & Wiekart, D. P. (2002). *Educating Young Children: Active Learning Practices for Preschool and Child Care Programs* (2nd Ed.) Ypsilanti, Michigan: High/Scope Press.

Mowbray, C.T., Holter, M.C., Teague, G.B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24 (3),* 315-340.

National Association for the Education of Young Children & National Council of Teachers of Mathematics. (2002). *Position statement. Early childhood mathematics: Promoting good beginnings.* Retrieved March 15, 2006, from http://www.naeyc.org/about/positions/pdf/psmath.pdf.

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

What Works Clearinghouse. (2004). Retrieved July 6, 2004, from www.w-w-c.org.