TEACHING MATH TO YOUNG CHILDREN PRACTICE GUIDE REVIEW PROTOCOL, VERSION 2.1

This protocol guided the review of research that informed the recommendations contained in the What Works Clearinghouse (WWC) practice guide "Teaching Math to Young Children," published in November 2013. The research review involved the following steps:

- The research staff searched the professional literature to identify relevant studies. Additional studies were identified by the expert panel.

- Studies were screened to determine whether they were within the scope of the practice guide.

- Eligible studies were assessed against WWC evidence standards.

    - Studies that met WWC evidence standards and were related to a recommendation within the guide were used to identify the strength of the evidence for each recommendation.

    - Studies that did not meet WWC evidence standards could be used to provide examples of practices. [Note: This differs from the procedures for WWC intervention reports, which report findings only for studies meeting WWC evidence standards.]

This document contains information about: (1) the purpose statement that guided the work of the panel and the research team; (2) procedures for conducting the literature search; (3) eligibility criteria for reviewing relevant studies; and (4) technical issues including attrition and group equivalence. Please refer to the *WWC Procedures and Standards Handbook (version 2.1)* for additional information.

Purpose Statement

The purpose of the practice guide is to provide evidence-based recommendations on developmentally appropriate techniques that early childhood educators can use to improve the early math skills of young children. All of the recommendations target program-level staff such as early childhood teachers, early childhood program directors, and kindergarten teachers.

To develop the recommendations, the panel first examined research on instructional strategies and practices that help improve children's early math skills and support teachers in choosing mathematical activities intentionally geared to students' understanding and emerging capabilities. While considering the research, the panel also considered questions such as:

1. How can teachers use materials, such as manipulatives and mathematical story books, to implement effective early math instruction strategies within their existing classroom curriculum?

2. How can teachers apply diverse pedagogical methods to teach skills that require rote memorization (for example, learning the count sequence) *and* abstract thought (for example, metacognitive skills for verbalizing how a problem is solved)?

3. How can teachers recognize opportunities to extend children's mathematical thinking?

4. How can early math instruction be implemented in whole class, small group, or one-on-one settings?

5. How can teachers use formative assessment to understand individual children's early math skills?

6. How can understanding children's mathematical misunderstandings be used to inform and differentiate instruction?

7. How does early childhood mathematics fit into a wider program focused on development of the "whole child"?

PROCEDURES FOR CONDUCTING THE LITERATURE SEARCH

The literature search involved a keyword search of multiple databases to identify studies examining the effectiveness of strategies for teaching math to young children.

Keyword Search

Primary Objective. The primary objective was to identify relevant practices for review by (1) identifying practices with potentially eligible studies, and (2) determining the approximate number of eligible studies related to each practice.

Search Strategy. Keywords were selected that aimed to capture literature related to teaching math to young children. Keywords related to outcomes, teaching, and age/grade levels were included to focus the search on literature that met the eligibility criteria for this review (see p. 5). The keyword list appears below. The list of databases that were searched appears in the next section.

- ***Keywords that identified teaching math:*** math*, number*, numeracy, counting.

- ***Keywords that identified relevant samples:*** "early childhood," pre-k, prekindergarten, pre-kindergarten, "early intervention," childcare, "child care," preschool, pre-school.

Multiple-word phrases listed above were searched as phrases. The keywords in each of the above categories were linked together with OR in a search so that they identified all articles that focused on any of the terms. The sets of search terms were then linked together with AND in a search so that they identified all articles that focused on teaching math to young children and

studied children in relevant programs. Finally, variations of words (e.g., "math*," to capture studies including the words "math" and "mathematics") were searched to ensure that our search was as inclusive as possible.

Databases

The search was conducted using the following databases:

- ***Academic Search Premier.*** This multidisciplinary database provides full text for more than 4,500 journals, including full text for more than 3,700 peer-reviewed titles. PDF backfiles to 1975 or further are available for well over 100 journals, and searchable cited references are provided for more than 1,000 titles.

- ***EconLit.*** EconLit, the American Economic Association's electronic database, is the world's foremost source of references to economic literature. The database contains more than 785,000 records from 1969 to the present. EconLit covers virtually every area related to economics.

- ***Education Research Complete.*** Education Research Complete is the definitive online resource for education research. Topics covered include all levels of education from early childhood to higher education, and all educational specialties, such as multilingual education, health education, and testing. Education Research Complete provides indexing and abstracts for more than 1,840 journals, as well as full text for more than 950 journals, and includes full text for more than 81 books and monographs, and for numerous education-related conference papers.

- ***EJS E-Journals.*** E-journals from EBSCO host® provide article-level access for thousands of e-journals available through EBSCO's Electronic Journal Service (EJS).

- ***ERIC.*** Funded by the U.S. Department of Education (ED), ERIC is a nationwide information network that acquires, catalogs, summarizes, and provides access to education information from all sources. All ED publications are included in its inventory.

- ***PsycINFO.*** PsycINFO contains more than 1.8 million citations and summaries of journal articles, book chapters, books, dissertations, and technical reports, all in the field of psychology. Journal coverage, which dates back to the 1800s, includes international material selected from more than 1,700 periodicals in more than 30 languages. More than 60,000 records are added each year.

- ***SocINDEX with Full Text.*** SocINDEX with Full Text is the world's most comprehensive and highest-quality sociology research database. The database features more than 1,986,000 records with subject headings from a 19,600+ term sociological thesaurus designed by subject experts and expert lexicographers. SocINDEX with Full Text contains full text for 708 journals dating back to 1908. This database also includes full text for more than 780 books and monographs and full text for 9,333 conference papers.

"Fugitive" or "Grey" Literature

"Fugitive" or "grey" literature refers to studies that are not published commercially or are otherwise inaccessible through conventional literature searches. To be considered by the WWC, these studies must be available to the public. To identify "fugitive" or "grey" literature for this review, the review team solicited recommendations from panel members.

ELIGIBILITY CRITERIA FOR REVIEWING RELEVANT STUDIES

Studies identified through the literature search were screened for relevance according to the eligibility criteria described in this section.

Populations to Be Included

Children must have been between ages 3 and 6, with the majority being under 6 (e.g., mean or median age of 5 yrs 6 months or younger) or in kindergarten when the practice or intervention was administered. Studies that contained children of other ages were not included unless (1) study results disaggregated the results by age, or (2) students of eligible ages represented over 50% of the aggregated mixed-age sample. Because the practice guide focused on foundational strategies for supporting early math learning, the panel did not distinguish between students with and without identified learning disabilities. Study samples could have been drawn from outside the United States, and practices and interventions could have been administered in any language.

Types of Practices and Interventions to Be Included

The guide considered studies of branded comprehensive or supplemental curricula or replicable strategies for teaching math to children attending preschool, prekindergarten, or kindergarten (or between the ages of 3 and 6 years). These may have included strategies or curricula used by teachers in classrooms, those used by math specialists in the school, those for use by paraprofessional educators or tutors, or those used by researchers working in the school.

Types of Research Studies to Be Included

To be included in the review, the study must have been written in English and have met the following relevancy criteria:

Topic relevance. The recommendations in the practice guide focused on instructional strategies to develop early math skills, including: (1) understanding concepts of numerical order (stable order principle), (2) one-to-one correspondence, (3) understanding the numerosity of a set (cardinality), (4) knowledge of the number sequence, (5) conservation principle—understanding that the count for a set of a group of objects stays the same no matter whether the objects are spread out or close together, (6) order irrelevance principle—understanding that the counting of objects can begin with any object in a set and the total will stay the same, (7) counting, (8) magnitude comparison, (9) basic number operations, including simple addition and subtraction,

(10) number sense, (11) recognizing numbers, (12) pattern recognition, (13) classification, and (14) size, measurement, and geometry. Research studies had to be directly related to one of these topics.

Time frame relevance. The study had to have been published between 1989 and 2011; earlier or later work may have been reviewed if suggested by a panelist. This time frame was established in order to define a realistic scope of work for the review. The time frame helps focus the review on interventions that may be available to teachers today. It also helps ensure that effectiveness is characterized relative to conditions similar to those in preschools, pre-kindergarten, and kindergarten programs today.

Study design relevance. Only empirical studies that used quantitative methods and inferential statistical analysis and that take the form of a randomized controlled trial (RCT) or used a quasi-experimental design (QED), a regression discontinuity design (RD), or a single-case design (SCD) were eligible for this review.

Intervention and comparison group relevance. Eligible intervention and comparison groups included:

- Intervention groups that received "bundled" interventions (that is, the intervention may have been multi-faceted and included multiple components)

- Multiple levels of intervention (for example, Intervention A might have been compared with Intervention A+B)

- Multiple comparison groups, typically other interventions (the guide prioritized the comparison most relevant for a recommendation but may have used each of the comparisons or combined groups where appropriate)

- Adjacent cohorts (for example, collection of data on an intervention group in one year and collection of data on a comparison group in the next year)

- Multiple cohorts (for example, an analysis of intervention vs. comparison in 2005, an analysis of different intervention and comparison groups in 2006); the guide reported an average of effects across cohorts

Types of Outcomes to Be Included

To be included in the review, the study must have examined the effect of the practice on an outcome related to early math skills. Eligible outcomes for this practice guide were classified into six domains:

- ***Basic number concepts*** included skills such as counting, magnitude, number line estimation, and finding counting mistakes.
- ***Number operations*** included age-appropriate addition or subtraction problems.

- ***Geometry*** included the identification of shapes and shape attributes, like angles and corners, as well as spatial relationships.
- ***Patterns and classification***included skills such as the child's ability to identify patterns in set of objects, duplicate patterns provided by another person, create and replicate demonstrated patterns, and sorting objects by similar and different characteristics.
- ***Number recognition*** included skills such as the child's ability to recognize any representation of a number (word, symbol, dots, etc.).
- ***General numeracy*** included outcome measures that covered two or more of the previous content areas.

Other information about or requirements for outcomes included the following:

Overalignment of outcomes. Outcome measures were considered overaligned with an intervention if the measure included some of the same materials that were used in the intervention or the measure was administered to the treatment group as part of the intervention. Outcome measures that were determined to be overaligned with an intervention were not included in determining the intervention's ratings.

Timing of outcome measurement. If more than one outcome meauure was reported, the outcome measurement closest to the end of the intervention was considered the primary outcome and labeled the "posttest." Subsequently measured outcomes were labeled "maintenance" outcomes. Multiple comparison adjustments were made when there was more than one posttest or maintenance outcome in the same domain.

Reliability. For RCTs and QEDs, the reliability of outcome measures (internal consistency, temporal stability/test-retest reliability, and inter-rater reliability) was assessed using the following WWC standards:

- Internal consistency: minimum of 0.60
- Temporal stability/test-retest reliability: minimum of 0.40
- Inter-rater reliability (percent agreement, correlation, Kappa): minimum of 0.50

If the reliability of each outcome measure was not specified in the study, data from the test or scale's publisher or other sources were used to establish the reliability of an outcome measure. If there was insufficient information, or the outcome measure did not meet the criteria above, the panel chair determined if the outcome measure should be considered reliable.

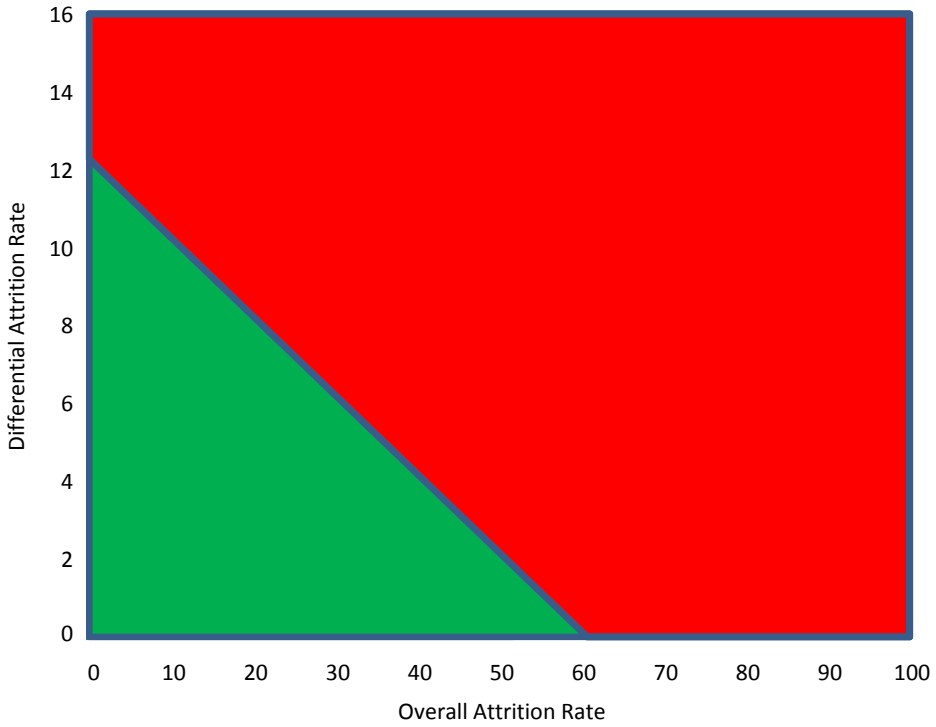
Statistical and Technical Issues

Eligible studies were assessed against WWC evidence standards, as described in the *WWC Handbook* and specified in this section.

Attrition in RCTs

As described in the *WWC Procedures and Standards Handbook (version 2.1),* the WWC is concerned about overall and differential attrition from the intervention and comparison groups for RCTs, as both contribute to the potential bias of the estimated effect of an intervention. The attrition bias model developed by the WWC was used in determining whether a study met WWC evidence standards (see Appendix A of the *Handbook*).

The review used the liberal (optimistic) threshold to assess attrition. This boundary selection was based on the assumption that most attrition in studies of early math was due to factors that were not strongly related to intervention status, such as parent mobility and absences on the days that assessments are conducted. When the combination of overall and differential attrition rates caused an RCT study to fall in the green area on the diagram shown below, the attrition was considered "low" and the level of bias acceptable. For RCTs with combinations of overall and differential attrition rates in the red area, the attrition was considered "high" and potentially had high levels of bias, and therefore, must have demonstrated baseline equivalence.



Group Equivalence in RCTs with High Attrition and QEDs

If the study design was a QED or an RCT with high levels of attrition, the study must have demonstrated baseline equivalence of the intervention and comparison groups for the analytic sample. The onus for demonstrating equivalence in these studies rested with the authors. Sufficient reporting of pre-intervention data must have been included in the study report to allow

the review team to draw conclusions about the equivalence of the intervention and comparison groups. For this review, the characteristic on which studies must have demonstrated equivalence was a pretest of the outcome measure (i.e., parallel form). If equivalence was demonstrated using a measure that was not a pretest, the evidence coordinator consulted with the panel chair on whether the measure was sufficient, with the determination based on considerations like the reliability of the measure and the relationship between the measure and the outcome measure. If baseline demographic characteristics (income, gender, race, special education, or English language learner status) were provided, reviewers calculated equivalence and reported in the study review guide. Panelists may have used this information when considering the evidence base, but demographic differences were not considered for the study rating.

Groups were considered equivalent if the reported differences in pre-intervention test scores were less than or equal to one-quarter of the pooled standard deviation in the sample, regardless of statistical significance. However, if differences were greater than 0.05 standard deviations and less than or equal to one-quarter of the pooled standard deviation in the sample, the analysis must have controlled analytically for the individual-level pre-intervention score(s) on which the groups differ (see Statistical and Analytical Issues below). If pre-intervention differences were greater than 0.25 for *any* of the listed scores in the same domain, the domain did not meet standards. In addition, if there was evidence that the populations were drawn from very different settings (such as rural vs. urban, or high-SES vs. low-SES), the chair or evidence coordinator may have decided that the environments were too dissimilar to provide an adequate comparison.

Statistical and Analytical Issues

Statistical controls. RCT studies with low attrition did not need to use statistical controls in their analyses, although statistical adjustment for well-implemented RCTs was permissible and could have helped generate more precise effect size estimates. For RCTs, the effect size estimates were adjusted for differences in pre-intervention characteristics at baseline (if available) using a difference-in-differences method if the authors did not adjust for pretest (see Appendix B of the *Handbook*). Beyond the pre-intervention characteristics required by the equivalence standard, statistical adjustment could have been made for other measures in the analysis as well, although they were not required.

This review preferred studies to report on and calculate effect sizes for post-intervention means adjusted for the pre-intervention measure. If a study reported both unadjusted and adjusted post-intervention means, the WWC review reported the adjusted means and unadjusted standard deviations.

Adjustments to statistical significance. The statistical significance of group differences were recalculated if (1) the study authors did not calculate statistical significance, (2) the study authors did not account for clustering when there is a mismatch between the unit of assignment and unit of analysis, or (3) the study authors did not account for multiple comparisons when appropriate. Otherwise, the review team accepted the calculations provided in the study.

When a misaligned analysis was reported (i.e., the unit of analysis in the study was not the same as the unit of assignment), the statistical significance of the effect sizes computed by the WWC incorporated an adjustment for clustering. The default intraclass correlation used for the students

was 0.20 for all outcomes. For an explanation of the clustering correction, see Appendix C of the *Handbook*.

When multiple comparisons were made within an outcome domain and not accounted for by the authors, the WWC accounted for this multiplicity by adjusting the reported statistical significance of the effect using the Benjamini-Hochberg procedure. Adjustments were made within a domain and within a particular intervention contrast. For example, if a study has five outcomes in a single domain and only two groups (Treatment 1 and Comparison), the denominator for the Benjamini-Hochberg procedure was five. Adjustments were made only for the number of outcomes that are eligible for review within a domain; if a study included three outcomes, but one outcome does not meet standards due to unacceptable reliability, then the WWC adjusted for the two outcomes within the domain that were reported.

Studies may also have multiple groups, and thus, more than one contrast of interest. For example, consider a study that has five outcomes in a single domain and three groups (Treatment 1, Treatment 2, and Comparison). If the guide was interested in both Treatment 1 vs. Comparison and Treatment 1 vs. Treatment 2, then the total number of comparisons to be adjusted is 10 (five from the Treatment 1 vs. Comparison contrast, and five from the Treatment 1 vs. Treatment 2 contrast). If the guide was interested in only Treatment 1 vs. Comparison, then the total number of comparisons to be adjusted is five.