MATHEMATICAL PROBLEM SOLVING 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 "Improving Mathematical Problem Solving in Grades 4 through 8," published in May 2012. 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 practice guide panel aimed to provide evidence-based recommendations to teachers about how to teach mathematical problem solving to students in grades 4 through 8. The panel examined research on instructional strategies that help students use problem-solving approaches to understand mathematical operations, reason strategically, and transfer knowledge to an applied context.

When considering the research and developing the recommendations, the panel considered questions such as:

1. How should teachers prepare to teach problem solving?
2. Should teachers provide students with explicit problem solving steps?
3. What instruction tools (e.g., visual representations) should teachers use?
4. Should teachers teach multiple problem solving strategies?
5. How should teachers connect problem solving to formal mathematics?

PROCEDURES FOR CONDUCTING THE LITERATURE SEARCH

The literature search involved a keyword search of multiple databases to identify effectiveness studies relevant to mathematical problem solving.

Keyword Search

Primary Objective. The primary object 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 mathematical problem solving. Keywords related to outcomes, teaching, and 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 word problems concepts:*** word problems, problem solving, problem-solving.

- ***Keywords that identified mathematical problem solving:*** math, algebra, prealgebra, arithmetic.

- ***Keywords that identified studies with outcomes:*** achievement, improvement, assessment, instructional effectiveness, effectiveness, outcome, skill.

- ***Keywords that identified interventions with teaching components:*** teach, strategy, instruction, curriculum, approach, monitor, intervention, training, self-regulation, metacognition, transfer.

- ***Keywords that identified relevant grade levels:*** K–8, K–3, K–6, elementary, middle school, kindergarten, first grade, first-grade, second grade, second-grade, third grade, third-grade, fourth grade, fourth-grade, fifth grade, fifth-grade, sixth grade, sixth-grade, seventh grade, seventh-grade, eighth grade, eighth-grade.[1]

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 five sets of search terms were then linked together with AND in a search so that they identified all articles that focused on mathematical problem solving, had relevant outcomes, included teaching, and studied relevant grades. Finally, variations of words (e.g., "achieve*," to capture studies including the words "achieve" and "achievement") were searched to ensure that our search was as inclusive as possible.

[1] Originally, the practice guide focused on students in grades K–8. The panel later decided to focus the guide on grades 4–8, and studies involving grades K–3 were excluded during review.

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.

- ***Campbell Collaboration.*** C2-SPECTR (Social, Psychological, Educational, and Criminological Trials Register) is a registry of more than 10,000 randomized and possibly randomized trials in education, social work and welfare, and criminal justice.

- ***Cochrane Central Register of Controlled Trials.*** Cochrane Central Register of Controlled Trials is a bibliography of controlled trials identified by contributors to the Cochrane Collaboration and others, as part of an international effort to hand search the world's journals and create an unbiased source of data for systematic reviews.

- ***Cochrane Database of Systematic Reviews.*** Cochrane Database of Systematic Reviews contains full-text articles, as well as protocols focusing on the effects of health care. Data are often combined statistically (with meta-analysis) to increase the power of the findings of numerous studies, each too small to produce reliable results individually.

- ***Cochrane Methodology Register.*** The Cochrane Methodology Register (CMR) is a bibliography of publications that report on methods used in the conduct of controlled trials. It includes journal articles, books, and conference proceedings that are taken from the MEDLINE database and from hand searches. The database contains studies of methods used in reviews and more general methodological studies that could be relevant to anyone preparing systematic reviews. CMR records contain the title of the article, information on where it was published (bibliographic details), and sometimes a summary of the article. CMR is produced by the UK Cochrane Centre, on behalf of the Cochrane Methodology Review Group.

- ***Database of Abstracts of Reviews of Effects.*** The Database of Abstracts of Reviews of Effects includes abstracts of published systematic reviews on the effects of health care from around the world, which have been critically analyzed according to a high standard of criteria. This database provides access to quality reviews in subjects for which a Cochrane review may not yet exist.

- ***Dissertation Abstracts.*** Dissertation Abstracts is a definitive subject, title, and author guide to virtually every US dissertation accepted at an accredited institution since 1861. Selected master's theses have been included since 1962. In addition, since 1988, the database includes citations for dissertations from 50 British universities that have been collected by and filmed at the British Document Supply Centre. Beginning with Volume 49, Number 2 (Spring 1988), citations and abstracts from Section C, Worldwide Dissertations (formerly European Dissertations) have been included in the file. Abstracts are included for doctoral records from July 1980 (Dissertation Abstracts International, Volume 41, Number 1) to the present. Abstracts are included for master's theses from spring 1988 (Masters Abstracts, Volume 26, Number 1) to the present.

- ***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.

- ***WorldCat.*** WorldCat is the world's largest network of library content and services. It allows users to simultaneously search the catalogs of more than 10,000 libraries, containing more than 1.2 billion books, dissertations, articles, CDs, and other media.

"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

Students must have been in grades 4–8 when the practice or intervention was administered. Studies that contained students in other grades were not included unless (1) study results disaggregated the results of students in eligible grades, or (2) students in eligible grades represented over 50% of the aggregated mixed-age sample. Studies that included students in grades 3 and 9 may have been reviewed if the panel believed findings were likely to be applicable to students in grades 4–8. 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 problem solving to students in fourth grade through eighth grade. These may have included strategies or curricula used by teachers in classrooms, those used by math specialists in the school, or those for use by paraprofessional educators, tutors, or parents.

Types of Research Studies to be Included

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

Topic relevance. The recommendations in the practice guide focused on instructional strategies to develop students' proficiency in mathematical problem solving, including: (1) understanding mathematical concepts and operations, (2) reasoning strategically, and (3) transferring knowledge to an applied context.

Time frame relevance. The study had to have been published between 1989 and October 2009; 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. Rigorous evaluations of practices and interventions implemented in the 20 years prior to the literature search test versions most likely to be available today and that were tested under conditions more likely to be similar to those existing 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), a single-case experimental design (SCD), or a strong correlational analysis were eligible for this review. For this review, the following analyses were considered strong correlational analyses: (1) an analytical model that included the pretest score as a statistical covariate (e.g., ANCOVA), (2) a student fixed-effects analytical model, and (3) a two-stage least-squares model that used a valid instrumental variable.

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

Eligible outcomes related to student achievement—specifically, problem solving in mathematics, as well as more general math achievement. Class grades and computation problems were not acceptable outcomes. Outcomes for this guide were classified by both mathematical competency domain and mathematical content area. The mathematical competency domains were:

- ***Procedural knowledge*** was defined as whether students correctly chose which mathematical operations and procedures helped them solve the problem, and how well they carried out the operations and procedures they chose to use. When students correctly solved a mathematics problem, they likely chose the appropriate operation or procedure and executed it correctly.

- ***Conceptual knowledge*** was defined as how well students understood mathematical operations and procedures, as well as the language of mathematics. One way for students to express their conceptual understanding of mathematics was by explaining the operations and procedures used to solve a problem accurately and completely.

- ***Procedural flexibility*** was defined as whether students can identify and carry out multiple methods to solve mathematics problems. If students solved a math problem in multiple ways, then they have likely developed procedural flexibility, a skill that may help them solve problems more efficiently in the future.

The mathematical content areas were:

- ***Problem solving focused on numbers and operations***—the content of the outcome measure involved using addition, subtraction, multiplication, division, and other procedures with whole numbers and rational numbers (integers, fractions, decimals, and ratios).

- ***Problem solving focused on algebra***—the content of the outcome measure involved using equations and symbols to represent unknown variables.

- ***Problem solving using geometry and measurement***—the content of the outcome measure involved the relations, properties, and measurements of solids, surfaces, lines, points, and angles, or consistently assigning numbers to phenomena.

- ***Problem solving focused on data analysis and probability***—the content of the outcome measure involved models, data transformation (including the use of graphs and charts), or statistics.

- ***General math achievement***—the content of the outcome measure covered two or more of the previous content areas.

Other information about or requirements for outcomes included the following:

Overalignment of outcomes. Outcome measures could have been 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. The outcome measurement closest to the end of the intervention was the primary outcome and labeled the "posttest." Subsequently measured outcomes were labeled "maintenance" outcomes. Outcomes that involved knowledge transfer between the intervention and measure were labeled "transfer" outcomes. Multiple comparison adjustments were made when there was more than one posttest, maintenance, or transfer outcome in the same mathematical competency 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. Ultimately, the panel chair made a determination.

SCD outcomes that involved written responses did not need to meet SCD inter-assessor agreement requirements if the evidence coordinator determined that the responses could be correctly scored by a single coder with a high degree of reliability. An example outcome that did not require reported inter-assessor agreement was a math test where students answered by writing numbers.

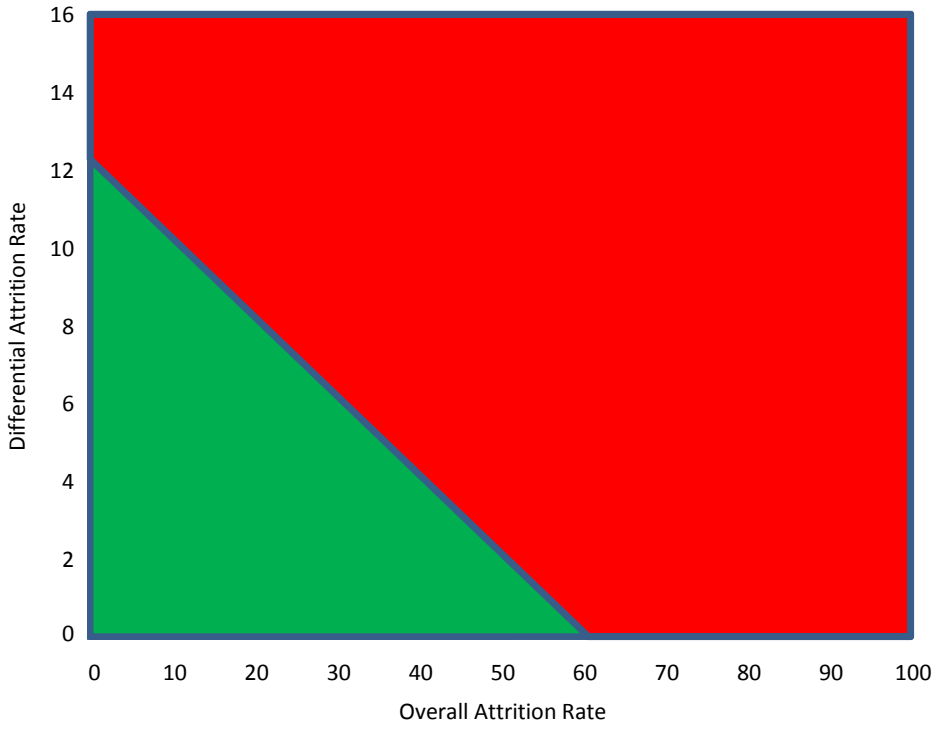
STATISTICAL AND TECHNICAL ISSUES

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*).

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. This boundary was based on the assumption that most attrition in studies of mathematical problem solving 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.



Group Equivalence in RCTs/RDs 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) or a standardized general math test (e.g., Iowa Test of Basic Skills). If equivalence was demonstrated using a measure that was not a pretest or a standardized test, 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 study 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. If adjusted post-intervention means were not reported, they were requested from the authors.

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 a mathematical competency 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 correction. If a study included more than two groups, when adjusting for multiple comparisons, reviewers counted the total number of outcome/pair combinations within a study relevant to a particular intervention. For example, in a study that has five outcomes in a domain, and three groups (Intervention, Comparison 1, Comparison 2) where all groups have data on all five outcomes, the total number of groups for a multiple comparison adjustment for Intervention will be ten (five outcomes in Intervention vs. Comparison 1 plus the five outcomes in Intervention vs. Comparison 2). If the study contains multiple age groups (some of which are outside the protocol) or multiple intervention groups, the evidence coordinator determined the correct adjustment procedure.