Research Paper

ROE

ECNU Review of Education
2021, Vol. 4(1) 65–83

DOI: 10.1177/2096531120944410
journals.sagepub.com/home/roe

SAGE

ECNUP

What Can China Learn From Evidence-Based Educational Reform? A Comparative Review of Educational Technology Programs' Effects on Mathematics Achievement

Chen Xie（谢晨）
Institute of International and Comparative Education, East China Normal University

Abstract

Purpose: This review contrasts the U.S. and Chinese mainland in terms of educational technology programs' effects on mathematics achievement. The existing evaluation studies are assessed and compared, especially from the perspective of research quality. Moreover, this review finds out what intervention programs work in these two countries and classifies evidence levels of effective programs.

Design/Approach/Methods: A systematic review and meta-analysis method is used.

Findings: The application of educational technology programs produced a better effect in Chinese mainland than in the U.S., however, the quantity and quality of evaluation studies is concerning in Chinese mainland. Since 2010, four programs in the U.S. and one program in Chinese mainland have qualified for strong evidence of effectiveness, which are worthy of receiving scale-up grants.

Corresponding author:
Chen Xie, Institute of International and Comparative Education, Faculty of Education, East China Normal University, Shanghai 200062, China.
Email: delxie1985@163.com

Originality/Value: This is the first systematic review to contrast the U.S. and China in terms of educational technology programs' effects on mathematics achievement. Determining what works through high-quality experiments is probably the most important requirement for evidence-based reform in education. This review compares the developments of the U.S., a leader of the movement, with those of China, an undecided observer. The review may provide enlightenment for the undecided countries and regions.

Keywords

Educational technology, evidence-based reform, mathematics achievement, meta-analysis

Date received: 9 April 2020; accepted: 2 July 2020

Developed countries including the U.S. and the UK have already made significant improvements in changing their educational policies to support the use of evidence in schools. Their aim is to put education into a virtuous cycle of innovation, evaluation, and dissemination like the transformed fields, such as medicine, agriculture, and technology (Haskins, 2014; Kolada, 2013).

The interest in evidence has grown in China, but it is far from prompting the construction of a research area of educational program development and evaluation. Initiatives similar to evidence-based education reform (Slavin, 2020) have not yet been followed by official recognition. Nevertheless, there are still a few experimental or quasi-experimental studies in China that are comparable with evaluation studies of program effectiveness in Western countries. Therefore, the present review is intended to systematically compare the development of evaluation studies in the U.S., a leader, with those in China, an undecided observer. Our findings may help China and similar countries determine where they are and what they want to do regarding the wave of evidence-based educational reform. Given that existing evaluation studies involve many areas, the present review takes the example of educational technology programs' effects on mathematics achievement, since educational technology application and mathematics education are topics that people have shown great interest in.

Evidence-based reform in education

Evidence-based reform in education refers to policies that enable educators and policymakers to use programs and practices proven to be effective in rigorous scientific studies (Bridgeland & Orszag, 2013; Gueron & Rolston, 2013; Slavin, 2017, 2020). This definition does not refer to simply disseminating generic principles of effective practices but to specific programs resulted to be effective and replicable through experimental evaluations (Slavin, 2020). Evidence-based reform consists of four essential activities generally (Slavin, 2013): (a) finding out what works in high-quality experiments, (b) making educators and policymakers aware of proven programs,

(c) providing incentives and resources for schools to implement proven programs, and (d) creating policies and systems to continually add to knowledge and effective practice.

For a very long time, educational policies were based on ideological standpoints or individual views; however, they did not always have the desired effects, as these policies were not combined with effective educational programs (Slavin, 2016). Programs with strong evidence will probably replicate positive outcomes if they are implemented as the way in the validating research. Advances in medicine, agriculture, technology, and many other fields have witnessed the success of evidence-based policies and practices (Haskins, 2014; Kolada, 2013).

Some developed countries have already made improvements in changing their policies to encourage the use of evidence in schools. Government initiatives in the U.S. and UK, such as Investing in Innovation, Education Innovation and Research, Institute for Education Sciences, and Education Endowment Foundation, have invested in the development and evaluation of many programs. Some official organizations and nongovernment institutes, like What Works Clearinghouse, Best-Evidence Encyclopedia, and Evidence for Every Student Succeeds Act (ESSA), are devoted to the interpretation of program evaluation studies, making policymakers and educators aware of proven programs. In the U.S., for example, the main federal education law, the 2015 ESSA offered incentives and financial support to schools to select and implement practices based on evidence of impact and included the definitions of different levels of evidence.

Effects of educational technology programs on mathematics achievement

Mathematics is one of the most important subjects in K-12 classrooms for almost all countries in the world. It is widely believed that students' performance in mathematics is of great importance to their nation's competitive strength for the future (Slavin et al., 2009). Many believe that secondary mathematics achievement is a key predictor of a nation's long-term economic potential. Therefore, whether in China, the U.S., or other countries, educators are always interested in knowing how to improve students' mathematics achievement. Researchers have developed and examined various programs to increase mathematics achievement (e.g., Pellegrini et al., 2018; Slavin & Lake, 2008), and one direction among them is to integrate educational technology into teaching and learning.

With the rapid development of information and communication technology, the number and citation of studies related to educational technology is booming. More than 20 systematic reviews and meta-analyses have been carried out to examine the effects of educational technology programs on mathematics achievement (Xie et al., 2020). The term computer-assisted instruction (CAI) was adopted in the early days of computer use in education, indicating either a type of software program for education or a type of instructional process (Dede, 2008). Rapid changes in technology lead to changing conceptualizations of the use of computers in educational practice.

In the 1980s, CAI focused on the use of mainframe-based computer applications and software in education, but at the very beginning of the 21th century, it moved to the effectiveness of microcomputer-based software on students. In the late 2000s, educational technology shifted its focus to computer software rather than computer hardware. Slavin et al. (2009) found that educational technology programs could be supplementary when students were sent to computer labs for additional, or they could be core, substantially replacing the teacher with self-paced instruction on the computer.

These reviews generally confirmed that educational technology had a positive influence on mathematics performance. However, the overall effect sizes in these meta-analyses varied considerably, from +0.10 to +0.54. Perhaps one of the main reasons for the variations in effect sizes was that each review used different inclusion criteria. For example, reviews carried out by Bangert-Drowns et al. (1985) and Kulik et al. (1985) selected studies using the following criteria: (a) the study should be carried out in an actual classroom setting; (b) the study should have a control group; (c) the attainment test of the study should be fair to the control group; and (d) the attrition rate of the study should be low. The reviews by Slavin and his colleagues (Cheung & Slavin, 2013; Pellegrini et al., 2018; Xie et al., 2020) adopted stricter inclusion criteria, requiring qualifying studies to show initial equivalence between the experimental and the control group, to have a minimum of 12 weeks' study duration, and a minimum of two teachers in each group to avoid the possible confounding of teacher effect with treatment effect. They also excluded overaligned and researcher-made measures, including studies in which students taught on computers were tested on computers which control students did not use. These rigorous standards could partly explain why their overall effect sizes were smaller than those of other reviews.

The overwhelming majority of evaluation studies in previous meta-analyses were conducted in English-speaking countries. Only a few meta-analyses have reviewed the effectiveness of educational technology programs in Chinese society (Liao, 1998, 2007; Xie et al., 2020).

Research purpose

To summarize, the current review contributes to the research field in several ways. To the best of my knowledge, this is the first systematic review to contrast the U.S. and China in terms of educational technology programs' effects on mathematics achievement. This review assesses and compares existing evaluation studies in these two countries, especially from the perspective of research quality. Moreover, this review finds out what intervention programs work in these two countries and classifies evidence levels of effective programs, which has different implications for future research, educational policy, and practice.

Determining what works through high-quality experiments is probably the most important requirement for evidence-based reform in education (Slavin, 2017, 2020). In this context, this review compares the developments of the U.S., a leader of the movement, with those of China, an undecided observer. The review is also the first attempt in the field and can provide enlightenment for other undecided countries and regions.

Method

This review comprised four key steps: (a) retrieving all potential studies, (b) screening studies by certain criteria, (c) coding qualified studies, and (d) computing effect sizes and implementing statistical analyses.

Literature search procedure

The document retrieval process consisted of several steps. First, we searched English databases, including Web of Science, Education Resources Information Center, JSTOR, PsycINFO, Education Full Text, ProQuest Dissertation & Theses, Digital Dissertation Consortium, and EdITLib. Three Chinese databases of the China National Knowledge Infrastructure were also retrieved: China Academic Journals Full-text Database (Core Journals), China Doctoral Dissertations Full-text Database, and China Masters' Theses Full-text Database. The index words consisted of educational technology, instructional technology, computer-assisted instruction, experiment, trial, intervention, and treatment. The time span was from 1960 to 2018.

Second, the previous meta-analyses and systematic reviews in the field were examined to see whether any studies were missed, especially Cheung and Slavin (2013), Pellegrini et al. (2018), and Xie et al. (2020), which have similar inclusion criteria with ours. We also followed up the references of all qualified studies to be sure that we do not miss any qualifying studies.

Criteria for inclusion

Based on the purpose of this review, we established the following inclusion criteria to identify possible qualifying studies.

1. The study topic was to assess the effects of educational technology programs on mathematics performance.
2. The study employed a pretest–posttest–control group design, in which the control group accepted traditional teaching and the intervention group used an educational technology program.
3. To ensure initial equality, the assignment of subjects was random or matched at pretest.
4. The study duration was no less than 12 weeks because we hope the studies are replicable in a realistic school context.

5. The study was conducted in Chinese mainland or the U.S. Studies that only focused on special groups, such as students with learning disability, were excluded.
6. The measuring tools of mathematics achievement should be quantitative and be fair to control groups. If the measurement centered on the topics which were only emphasized in treatment groups, the studies were excluded.
7. The study result should report effect sizes or include available data to calculate effect sizes.

Coding

The important study features were as follows: grade levels, year of publication, group design, outcome measure, sample size, and types of publication. The study features were sorted in the following ways:

1. Grade levels: elementary (Grades 1–6 in China or K-5 in the U.S.) and secondary (Grade 7–12 in China or 6–12 in the U.S.);
2. Year of publication: 1960s, 1970s, 1980s, 1990s, 2000s, and 2010s;
3. Group design: randomized controlled trials (RCTs) and quasi-experimental designs (QEDs);
4. Outcome measure: standardized tests and researcher-made measures;
5. Sample size (students): 1–200, 201–500, 501–1,000, and >1,000;
6. Types of publication: published and unpublished.

Effect size computation and statistical analyses

In this review, effect sizes referred to the standardized difference between experimental and control group posttests after adjustment for pretests and other covariates. We employed Comprehensive Meta-Analysis (V3) software to implement all statistical analyses. In computing the overall effect size, there are usually two statistical models, the fixed-effect model and the random-effect model. The former assumes that the studies included in the analysis are homogenous and the differences in observed effect sizes are attributed to sampling error; the latter, by contrast, assumes that the included studies are not identical functionally, and we should therefore not assume that they share a common effect (Borenstein et al., 2009; Schmidt et al., 2009). In this article, we employed both models to obtain the overall effect, but we maintained the random model was more suitable for our study. The reasons were that the studies included in this meta-analysis had some substantial differences and that the overall effect size could be generalized to a range of scenarios. Additionally, we used a heterogeneity test (Q-test) to show whether the true effect sizes varied from study to study. The Z-value was also calculated to test whether the true overall effect size was zero.

Table 1. Overall effect sizes.

					95% Confidence interval		Test of mean		Test of heterogeneity in effect sizes		
		k	ES	*SE*	Lower	Upper	*Z*-value	*p* Value	*Q*-value	*df*(*Q*)	*p* Value
The U.S.											
(1)	Fixed	78	.09	.01	.07	.10	12.03	.00	259.21	77	.00
(2)	Random	78	.13	.02	.10	.16	8.04	.00			
Chinese mainland											
(1)	Fixed	36	.25	.02	.21	.29	13.54	.00	114.39	35	.00
(2)	Random	36	.39	.04	.31	.48	9.00	.00			

Note. ES = Effect Size.

Results

Qualified evaluation studies

General characteristics. In the U.S., 78 studies met the inclusion criteria, covering a total sample size of 81,189 students in K-12. The overall effect size of these studies is +0.13 in the random effects model, and the *Z*-test demonstrated that the true effect was significantly larger than zero (see Table 1). The random effects model is considered more appropriate in the present review, since it is assumed that the populations represented by included studies are different in many features (e.g., types of intervention or duration of intervention). Moreover, the *Q*-test here supported our hypothesis, which indicated that there was a substantial variation in this collective set of studies.

In Chinese mainland, 36 qualified studies were included, with a total sample size of 13,438 students in Grades 1–12. The overall effect size is +0.39 in the random effects model, and the *Z*-test also demonstrated that the true effect was significantly different from zero (see Table 1). The result of the *Q*-test was also significantly heterogeneous.

The time span retrieved for this review is from 1960 to 2018, but the earliest included study in the U.S. was published in 1980. There are 21 studies published in the 1980s, 15 studies published in the 1990s, 26 studies published in the 2000s, and 16 studies published in the 2010s in the U.S. In terms of Chinese mainland, the earliest included study was published in 1999. Only 1 study was published in the 1990s, 16 studies were published in the 2000s, and 19 studies were published in the 2010s.

With regard to grade levels of student samples, 50 studies included participants in elementary schools, 26 studies included participants in secondary schools, and 2 studies included participants from both levels in the U.S. For studies conducted in Chinese mainland, 8 included participants in elementary schools and 28 included participants in secondary schools.

Characteristics of research quality. Although the present review employs strict inclusion criteria to exclude low-quality studies, this does not mean that the research quality of all included studies is the same. The included studies vary considerably in some important characteristics, which can affect the validity and reliability of the research findings and therefore affect the extent to which policymakers may utilize the research findings. The following features of the included studies are used by this review to evaluate research quality: group designs, outcome measures, sample size, and peer review.

The group designs of evaluation studies involve two categories in the current review: RCTs and QEDs. RCTs randomly assign students, classes, or schools to intervention groups and control groups, and the unit of analysis is at the same level of the random assignment. QEDs match intervention groups and control groups on key prior variables. Although this review requires QEDs to demonstrate baseline equivalence between different condition groups on measured characteristics, there may be differences in unmeasured characteristics that could introduce bias into an estimate of the effect of the intervention (What Works Clearinghouse, 2017). Bias is a systematic difference between the true impact of the intervention and the estimated impact, which can lead to incorrect conclusions about the effect of the intervention. For this reason, QEDs cannot receive the highest rating. The current review includes 31 RCTs and 47 QEDs in the U.S. However, among studies conducted in Chinese mainland, only 4 are RCTs and 32 are QEDs.

This review classifies outcome measures of included studies into two categories: standardized tests and researcher-made measures. Standardized tests are designed in such a way that the questions, conditions for administering, scoring procedures, and interpretations are consistent and are administered and scored in a predetermined, standard manner (Olson & Sabers, 2008). Standardized tests, in which the same test is given in the same manner to all test takers, are assumed to have face validity and be reliable. For researcher-made measures, this review rejects measures if they contain content provided to subjects in one condition, but not the other. Notwithstanding, Cheung and Slavin (2016) found that effect sizes for studies using experimenter-made measures were twice the size of effect sizes from standardized tests, which supports that students in intervention groups are usually more sensitive to experimenter-made measures than those in control groups. Unlike standardized tests, researcher-made measures are likely to lack enough evidence for face validity, reliability, or the consistency of outcome collection manner. Among 78 included studies from the U.S., 75 adopted standardized tests as outcome measures, and 3 used researcher-made measures. Among 36 studies from Chinese mainland, 1 used researcher-made measures, and standardized tests were employed by 35 studies in which 2 used the Trends in International Mathematics and Science Study, 31 used city/district/school-wide tests, and 2 used external experts-made measures. Only three U.S. studies employed district-wide tests.

Table 2. Sample size and mean effect sizes.

Sample size	Number of included studies in the U.S.	Mean effect size in the U.S.	Number of included studies in Chinese mainland	Mean effect size in Chinese mainland
0–200	24	.22	30	.44
201–500	20	.21	2	.44
501–1,000	14	.08	1	.40
1,001–14,000	20	.08	3	.17

In general, studies with a small sample size produce larger effect sizes than those with a large sample size (Cheung & Slavin, 2016), which can be called "super-realization effect" of small-scale studies. There are several possible explanations for the finding. First, researchers are more likely to sustain high implementation fidelity in a small-scale study or provide additional support that could never be replicated on a large scale. Second, small-scale studies may appear to have large effect sizes as their limited statistical power requires large effect sizes to reach statistical significance. If a small study happens to produce a large effect size, it is likely to be submitted and accepted somewhere and may therefore be easier for reviewers to find. According to sample sizes of included studies, Table 2 classifies them into four categories and shows the number of studies and the average effect size for each category. The number ($N = 24$) of small-scale studies (fewer than 200) in the U.S. is close to that ($N = 30$) in Chinese mainland, but the number ($N = 54$) of large-scale studies (more than 200) in the U.S. is much bigger than that ($N = 6$) in Chinese mainland.

Some included studies were published in peer review journals, whereas other included studies referred to as gray literature, or unpublished studies, including dissertations, working papers (technical reports), conference papers, and so on. Some meta-analysts argue that the quality of published articles is generally higher than that of unpublished studies, since published works have to go through a rigorous and double-blind peer review process. Although the present review used rigorous criteria to guarantee the quality of included studies, our criteria cannot cover everything. The peer review process may guarantee research quality in some way where this review missed. Among the included studies conducted in the U.S., there are 23 articles published in peer review journals, 27 doctoral dissertations, 1 master's thesis, 26 working papers, and 1 conference paper. Six China's studies were published in peer review journals, 29 were master's theses, and 1 was a working paper.

Effective programs

One objective of the systematic review of evaluation studies is to determine what works in the field, that is, which educational technology programs can effectively increase students'

mathematics achievement. As noted in the last section, the research quality of the included evaluation studies is different, and therefore, the evidence levels of program effectiveness they provide are different. The ESSA, which replaced the No Child Left Behind Act as the main U.S. education law in 2015, defined strong, moderate, and promising levels of evidence supporting educational programs.

The Evidence for ESSA (2017) developed detailed standards for these three categories as follows. Strong: at least one randomized, well-conducted study showing significant positive student outcomes, and no studies showing significant negative outcomes. Moderate: at least one quasi-experimental (i.e., matched), well-conducted study showing significant positive student outcomes, and no studies showing significant negative outcomes. Promising: Programs that would have qualified for "strong" or "moderate" but did not qualify because they failed to account for clustering (but did obtain significant positive outcomes at the student level) may qualify for this category.

According to these standards, the educational technology programs included in this review were classified into different evidence levels (see Table 3). Educational technology has been developing rapidly. The programs evaluated by older studies are probably out of date and inapplicable to the current situation. Hence, this section only includes programs that have been evaluated by studies released after 2010.

In the U.S., four educational technology programs have shown strong evidence of effectiveness on mathematics achievement: enhancing Missouri's Instructional Networked Teaching Strategies (eMINTS), ASSISTments, Carnegie Learning High School Math Solution, and DreamBox Learning.

eMINTS is a professional development approach designed to promote inquiry-based learning, high-quality lesson design, and technology-rich learning environments, as well as to build community among teachers and students (see Meyers et al., 2016, in Supplementary Material). Classes have interactive whiteboards, LCD projectors, and a 1:1 ratio of students to computers. Professional development is provided for principals, technology coordinators, and classroom teachers in middle schools to design high-quality inquiry-based lesson plans, implement inquiry-based learning strategies, build community among teachers and students, and integrate technology into classroom teaching. The model also has a strong emphasis on helping principals understand and lead the program implementation and carry out teacher observations to support teachers' instructional performance.

ASSISTments is an online homework tool coupled with teacher training, providing students with math assistance and teachers with assessments of student progress (see Roschelle et al., 2016, in Supplementary Material). There are two types of ASSISTment content: The first is linked to existing textbook and homework problems and provides teachers the flexibility to assign suitable

Table 3. Effective programs.

Program	Evidence level	Mean effect size	Evaluation studies
U.S.			
eMINTS	Strong	.15	Meyers et al. (2016)
ASSISTments	Strong	.18	Roschelle et al. (2016)
Carnegie Learning High School Math Solution	Strong	.03	Pane et al. (2014)
DreamBox Learning	Strong	.11	Wang and Woodworth (2011)
Time to Know	Promising	.31	Rosen and BeckHill (2012)
ST Math	Promising	.08	Rutherford et al. (2014)
Chinese mainland			
REAP	Strong	.18	Lai et al. (2015), Mo et al. (2013), Mo, Zhang, Luo, et al. (2014), and Mo, Zhang, Wang, et al. (2014)
Computer-assisted Inquiry Learning	Moderate	.44	Gao (2013), Liu (2012), Yuan (2012), and Zheng (2014)
Sketchpad-assisted Instruction	Moderate	.42	He (2015) and Liu (2013)
Cooperative Learning Assisted by Blogs and Concept Maps	Moderate	.61	Dai (2011)
Pad-based Instruction	Moderate	.38	Zhang (2017)
Developmental Homework	Moderate	.36	Cheng (2011)

Note. For complete details of references under "Evaluation studies," see Supplementary Material. eMINTS = enhancing Missouri's Instructional Networked Teaching Strategies; REAP = Rural Education Action Program; ST Math = Spatial Temporal Math.

problem sets; the second is intended for mastery-oriented skill practice. In its successful evaluation, the intervention included professional development and coaching designed to increase teachers' readiness and efficacy in using ASSISTments. The ASSISTments technology is a web-based platform available at no cost to schools; however, students need computer access at home.

Carnegie Learning High School Math Solution provides a first-year algebra course designed for students ranging in ability and grade levels (see Pane et al., 2014, in Supplementary Material). The curriculum combines traditional textbook and workbook materials with self-paced individualized instruction via automated tutoring software. MATHia, the technology component of Carnegie Learning High School Math Solution, is a web technology that runs on desktop/laptop computers as well as Chromebooks and iPad, Android, and Windows tablets.

DreamBox Learning is an adaptive, online math program designed to complement classroom instruction (see Wang & Woodworth, 2011, in Supplementary Material). The program is available in Grades K-8, but the only qualifying study involved Grades K-1. Combining a motivating, game-like environment with a rigorous, standards-aligned curriculum, it responds to learners' actions and decisions by continuously adapting to supporting student competency with math concepts and promoting strategies for fluency and application.

Two U.S. programs meet the standards of the "promising" level: Time to Know and Spatial Temporal Math (ST Math). Time to Know is a blended approach in which students use one-to-one laptops with interactive curriculum (see Rosen & BeckHill, 2012, in Supplementary Material). In math, teachers open each lesson with digital animations, followed by class discussion. Students work on their laptops to perform guided experiments, which they then post in a class digital gallery to be discussed by the class. Teachers receive extensive professional development and on-site coaching to integrate their teaching with the digital tools and modify them as needed.

ST Math is a supplemental online approach that provides students with a yearlong curriculum of games featuring visual interactive animated puzzles (see Rutherford et al., 2014, in Supplementary Material). ST Math progress requires 60–90 min of usage each week. Technology requirements are any modern device and O/S from desktops to Chromebooks to tablets (except smartphones whose screen sizes are too small for interactivity).

In Chinese mainland, the computer-assisted math remedial tutoring developed by the Rural Education Action Program (REAP) is the only program with strong evidence of effectiveness (see Lai et al., 2015; Mo et al., 2013; Mo, Zhang, Luo, et al., 2014; Mo, Zhang, Wang, et al., 2014, in Supplementary Material). The students in the treatment group had two 40-min sessions per week during lunch breaks or after school, which served as a supplement to the regular in-school instruction hours. The sessions were mandatory, and attendance was taken by the teacher-supervisors. During each session, two students shared one computer and played animation-based math games designed to help students review and practice the basic math material that was being taught in their regular school math classes. The material was remedial in nature, based on the material that was in student textbooks and material taught in the same week.

Five programs showed moderate evidence of effectiveness: computer-assisted inquiry learning, sketchpad-assisted instruction, cooperative learning assisted by blogs and concept maps, pad-based instruction, and developmental homework.

In the model of computer-assisted inquiry learning, a teacher usually creates a puzzling situation and identifies problems for inquiry and then students adopt educational applications to perform experiments to solve the problems (see Gao, 2013; Liu, 2012; Yuan, 2012; Zheng, 2014, in Supplementary Material). Students can share and discuss their ideas and findings with each other as well as a teacher. Finally, students summarize their inquiry process and obtain

mathematical knowledge with the help of a teacher. Frequently used educational applications include Geometer's Sketchpad, interactive whiteboards, and multimedia courseware.

Sketchpad-assisted instruction can employ Super's Sketchpad (see He, 2015, in Supplementary Material) or Geometer's Sketchpad (see Liu, 2013, in Supplementary Material) to raise the efficiency of transmission instruction. Super's Sketchpad and Geometer's Sketchpad are similar interactive software. They allow the creation of numerous figures that can be measured. The application is useful for showing the transformation and motion of figures and solving difficult mathematical problems.

Dai (2011, in Supplementary Material) integrated cooperative learning with blogs and concept maps as the intervention program. In the model, a teacher made the trunk of a concept map before class and taught the knowledge in the concept map in class; in teams, students developed and filled out the concept map after class and showed their products in blogs. Each team learned from and commented on other teams' concept maps, and each team took advantage of the comments from the teacher and other teams to improve its concept map. The teacher made a final evaluation of each team's learning process and products.

Pad-based instruction requires the provision of a pad for each student and teacher (see Zhang, 2017, in Supplementary Material). Students watch a 5-min video and complete the corresponding homework on the pad before class, and teachers can receive the data related to the students' homework. In class, students can see the learning content and interact with their classmates and the teacher on the pad. After class, students and parents can play back teaching videos.

Developmental homework can be run on the Moodle course platform (see Cheng, 2011, in Supplementary Material). Developmental homework involves five types of homework: preview, review, expansion, appreciation, and oral communication. The primary functions of the Moodle course platform include sharing learning materials, assigning online homework, and presenting excellent work.

Discussion and conclusion

China's Education Modernization 2035 proposes that one of the four essential routes of education modernization is first conducting trials in a pilot area and then pushing forward the reform systematically (Ministry of Education of the People's Republic of China, 2019). One question that naturally follows is, how can the trials be evaluated scientifically? The movement of evidence-based educational reform provides a model: Evidence of effectiveness has to be from rigorous experiments in which students engaged in innovative programs are compared over significant periods of time to students using traditional methods in terms of their gains on valid measures of achievement (Slavin, 2013, 2017, 2020). Based on this, evidence-based reform is dedicated to bringing education into a cycle of innovation, evaluation, and dissemination that has been proven

successful in fields such as medicine, agriculture, and technology (Haskins, 2014; Kolada, 2013). Currently, medical experts around the world are doing everything to fight the COVID-19. When evaluating whether a treatment is effective, what they rely on are not small trials with different methodologies but large, well-conducted RCTs (e.g., Ghebreyesus, 2020). Medical experts and their research approaches and methods are trusted by the vast majority of governments worldwide, as they have helped humanity overcome countless difficulties.

The most important requirement in evidence-based reform in education is to determine what works through high-quality experiments. In this respect, however, the research basis in Chinese mainland is extremely weak. As the leader of evidence-based reform in education, the U.S. is the most experienced country in this matter. It has produced systematic procedures and standards and achieved great progress. Therefore, using the U.S. as a benchmark, the present review is intended to give Chinese researchers, policymakers, and educators an understanding of where we are. By contrast, this review also aims to determine China's unique advantages and potential.

Qualified studies and overall effects

This review takes the example of educational technology programs' effects on mathematics achievement to make a comparison because integrating educational technology into the teaching process has recently become a highly popular approach, and mathematics is one of the most important subjects in K-12 education. Studies that examined the effectiveness of educational technology programs are retrieved comprehensively and screened using rigorous inclusion and exclusion criteria. The number of included studies in Chinese mainland is 46% of that in the U.S. (36 vs. 78). In fact, there were 36 studies meeting the inclusion criteria in the U.S. before 2000. The earliest of the included studies in the U.S. was released in 1980; however, in Chinese mainland, the earliest study was released in 1999, nearly 20 years later. In this sense, Chinese mainland's evaluation studies in the field are behind those in the U.S. by nearly 20 years.

The overall effect size ($d = +.39$) of the included evaluation studies in Chinese mainland is larger than that ($d = +.13$) in the U.S. This finding suggests the possibility that the application of educational technology programs can produce a better effect in Chinese mainland than in the U.S., which may stimulate more research in China in future. However, there are also competing explanations, such as differences in group designs, sample sizes, and outcome measures, as shown earlier in the "Results" section.

Quality of evaluation studies

In contrast to the difference in the number of evaluation studies, the research quality of studies in Chinese mainland is much more concerning. RCTs are the gold standard for estimating program effects. Only programs that are proven effective by RCTs qualify for the strong level of evidence in

the U.S. federal education law, the ESSA. However, the present review found only four RCTs in Chinese mainland, which is equivalent to 13% of the RCTs in the U.S. Although QEDs match different condition groups in mathematics achievement during pretests, they do not exclude the effects of unmeasured characteristics that could introduce bias into an estimate of program effects. Therefore, the findings from QEDs have to be examined by large-scale RCTs before the corresponding intervention programs can be applied on a large scale.

In terms of sample size, small-scale studies (no more than 200 participants) account for 83% of the included studies in Chinese mainland, and the sample sizes of only six studies were larger than 200 participants. In contrast, large-scale studies account for 69% of U.S. evaluation studies, and the sample sizes of 21 studies were larger than 1,000 participants. The overall effect of small-scale studies in Chinese mainland is up to +0.44, but this is probably a super-realization effect. It is extremely difficult to replicate the result in a large-scale application, as the implementers are less likely to sustain high implementation fidelity as researchers do in a small study. Furthermore, small studies with small effect sizes are infrequently published and are therefore infrequently retrieved by reviewers. In addition, 29 studies involved only two classes from the same school as participants in Chinese mainland. The external validity of these studies is therefore probably questionable. If these programs are applied by different teachers or in different schools, their effectiveness needs to be examined further.

With regard to outcome measures, city/district/school-wide tests and external expert-made measures can be considered standardized tests in the broad sense (see Lai et al., 2015; Mo, Zhang, Wang, et al., 2014, in Supplementary Material), but it must be noted that these tests are deficient in norm-reference and test equating (Wen, 2014; Xu & Wang, 2004). Such measurement problems may have an impact on their research findings.

Only six included studies were published in peer review journals in Chinese mainland. Three of them were published in Social Sciences Citation Index (SSCI) journals, and the other three were published in Chinese SSCI journals. In addition to the lack of researchers in the field, another possible reason for this small number of peer review studies is that academic journals in Chinese mainland do not pay much attention to evaluation studies on intervention programs. In contrast, these kinds of studies are highly popular in top SSCI journals related to educational research, such as *American Education Research Journal, Computers & Education, Educational Evaluation and Policy Analysis, Journal of Experimental Education*, and *Journal of Research on Educational Effectiveness*. On the other hand, publication bias is a common problem in both China and the U.S. Studies with small effect sizes or without significant results are less likely to be published, which makes readers overestimate the true effects and renders them unaware of which programs are ineffective.

The above findings have some implications for educational policy and research in Chinese mainland. First, more high-quality experimental studies, especially large-scale RCTs, are urgently

needed. This is the very first step in establishing educational policy and practices based on evidence. The government will play a key role in achieving this goal. Policy and funding should be skewed toward the research field, which will attract more researchers to engage in related studies and therefore create a strong and enterprising research community. Second, taking the *Gaokao* (national college entrance examination) reform as an opportunity, educational administrators and researchers should accelerate the development process of standardized tests that can reach the top level in the world. In the subjects and areas in which standardized tests are not currently applicable, the quality of educational testing needs to be improved. Third, academic journals in Chinese mainland need to be aware of the importance of program evaluation studies and provide more opportunity for them to be published. Governments or third-party organizations need to construct a database to help studies with small effect sizes or without significant results to report their findings publicly. Finally, given the development level of the field in Chinese mainland, international communication and cooperation should be encouraged and heavily sponsored, which will help local researchers catch up as soon as possible.

Effective programs

The high-quality evaluation studies included in the present review can help us determine what educational technology programs succeed in increasing mathematics achievement and what programs do not. Nevertheless, the evidence levels for these effective programs are different, and therefore, their implications for future research, policy, and practice are different. The Evidence for ESSA (2017) developed detailed standards to distinguish evidence levels of program effectiveness. The programs that already have strong evidence can qualify for scale-up grants. The programs with moderate evidence or promising evidence should be subsidized to further examine their effects using more rigorous methods. Schools or districts should first look for programs supported by strong evidence in a similar setting and/or population to the ones the schools or districts themselves serve. When strong evidence is not available, they can choose programs with moderate evidence. When strong evidence or moderate evidence is not available, the existence of promising evidence may suggest that a program is worth exploring.

In Chinese mainland, the only program that already has strong evidence is the remedial tutoring program developed by REAP (see Lai et al., 2015; Mo et al., 2013; Mo, Zhang, Luo, et al., 2014; Mo, Zhang, Wang, et al., 2014, in Supplementary Material). The program keeps the regular curriculum intact and is implemented outside of regular classes, such as during the midday rest or after school. Moreover, the program does not rely on schoolteachers, so it does not require investment in teacher training and may be cost-effective. Hence, the program is worthy of receiving scale-up grants.

In addition, there have been five programs meeting the standard of moderate evidence in China since 2010. However, we have to interpret their results with caution, as their evaluation studies only involved two classes of students. The external validity of these studies needs to be examined in a large and multisite sample. Furthermore, these programs have neither proper name nor specialized organization in charge of development and promotion. Therefore, they still have lots of work to do before they can be replicated.

Four programs have qualified for strong evidence of effectiveness in the U.S. since 2010. The description for these programs is shown in the "Results" section above, and more information can be found in the original studies. A common feature of eMINTS and ASSISTments is the inclusion of professional development and the training of teachers and school staff (Meyers et al., 2016; Roschelle et al., 2016). Both Carnegie Learning High School Math Solution and DreamBox Learning apply self-adaptation design to meet individualized learning needs (see Pane et al., 2014; Wang & Woodworth, 2011, in Supplementary Material).

When applying the external research findings to China's schools, we have to pay attention to Slavin's statement on evidence-based reform in education: Simply disseminating information about generic principles of effective practice has not generally been found to make much of a difference in practice; this reform is dedicated to promoting the use of specific programs found to be effective and replicable (Slavin, 2020). Therefore, it is a specific program rather than one feature or multiple features of effective programs that should be first applied and examined in China's schools. Given the intercultural differences, research teams may adjust a certain part of an intervention program to the needs and setting of participating schools and students, and even programs with strong evidence also need high-quality experiments to evaluate their effects on Chinese students. Some features of effective programs are likely to inspire researchers to create new programs, but these new programs cannot be considered to have an evidence base, as previous evaluation studies support the effectiveness of a certain combination of some elements.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funded by the National Social Science Fund of China (Education, 2019) under the program "Evidence-Based Evaluation of Educational Intervention Programs" (grant number CGA190249).

Supplemental Material

Supplemental material for this article is available online (https://journals.sagepub.com/doi/suppl/10.1177/2096531120944410).

References

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1985). Effectiveness of computer-based education in secondary schools. *Journal of Computer-Based Instruction*, *12*(3), 59–68.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to metaanalysis*. John Wiley.

Bridgeland, J., & Orszag, P. (2013). *Can government play moneyball?* http://www.theatlantic.com/magazine/archive/2013/07/can-government-play-moneyball/309389/

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

Cheung, A. C., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88–113. https://doi.org/10.1016/j.edurev.2013.01.001

Dede, C. (2008). Theoretical perspectives influencing the use of information technology in teaching and learning. In J. Voogt & G. Knezek (Eds.), *International handbook of information technology in primary and secondary education* (pp. 43–62). Springer.

Evidence for ESSA. (2017). *Evidence for ESSA standards and procedures*. http://www.evidenceforessa.org/

Ghebreyesus, T. A. (2020, March 18). *WHO director-general's opening remarks at the media briefing on COVID-19*. World Health Organization. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—18-march-2020

Gueron, J. M., & Rolston, H. (2013). *Fighting for reliable evidence*. Russell Sage.

Haskins, R. (2014, December 31). Social programs that work. *The New York Times*. http://www.nytimes.com/2015/01/01/opinion/social-programs-that-work.html?_r=0

Kolada, G. (2013, September 2). Guesses and hype give way to data in study of education. *The New York Times*. http://www.nytimes.com/2013/09/03/science/applying-new-rigor-in-studying-education.html?_rD0

Kulik, J. A., Kulik, C. L. C., & Bangert-Drowns, R. L. (1985). Effectiveness of computer-based education in elementary schools. *Computers in Human Behavior*, *1*(1), 59–74. https://doi.org/10.1016/0747-5632(85)90007-X

Liao, Y.-K. C. (1998). Effects of hypermedia versus traditional instruction on students' achievement: A meta-analysis. *Journal of Research on Computing in Education*, *30*(4), 341–359.

Liao, Y.-K. C. (2007). Effects of computer-assisted instruction on students' achievement in Taiwan: A meta-analysis. *Computers & Education*, *48*(2), 216–233. https://doi.org/10.1016/j.compedu.2004.12.005

Olson, A. M., & Sabers, D. (2008). Standardized tests. *21st Century Education: A Reference Handbook*, *1*, 423–430.

Ministry of Education of the People's Republic of China. (2019). *China's education modernization 2035* [in Chinese]. http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/201902/t20190223_370857.html

Pellegrini, M., Lake, C., Inns, A., & Slavin, R. E. (2018). *Effective programs in elementary mathematics: A best-evidence synthesis*. Center for Research and Reform in Education, Johns Hopkins University.

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed-versus random-effects models in metaanalysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 97–128. https://doi.org/10.1348/000711007X255327

Slavin, R. E. (2013). Overcoming the four barriers to evidence-based education. *Education Week*, *32*(29), 24.

Slavin, R. E. (2016). Getting to scale: Evidence, professionalism, and community. *Journal of Education for Students Placed at Risk (JESPAR)*, *21*(1), 60–63. https://doi.org/10.1080/10824669.2015.1110028

Slavin, R. E. (2017). Evidence-based reform in education. *Journal of Education for Students Placed at Risk (JESPAR)*, *22*(3), 178–184. https://doi.org/10.1080/10824669.2017.1334560

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, *55*(1), 21–31. https://doi.org/10.1080/00461520.2019.1611432

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, *78*(3), 427–515. https://doi.org/10.3102/0034654308317473

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, *79*(2), 839–911. https://doi.org/10.3102/0034654308330968

Wen, Z. (2014). Gaokao reform: Policy fairness and technical compatibility [in Chinese]. *Global Education*, *43*(2), 4–14.

What Works Clearinghouse. (2017). *Standards handbook (version 4.0)*. Institute of Education Sciences, U.S. Department of Education.

Xie, C., Cheung, A., Lau, W., & Slavin, R. E. (2020). The effects of computer-assisted instruction on mathematics achievement in mainland China: A meta-analysis. *International Journal of Educational Research*. https://doi.org/10.1016/j.ijer.2020.101565

Xu, J., & Wang, R. (2004). Communalities and differences in the understanding of standardized test [in Chinese]. *China Examinations*, *1*(1), 24–26.