education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 25 Number 5 January 23, 2017 ISSN 1068-2341

The Failure of the U.S. Education Research Establishment to Identify Effective Practices: Beware *Effective Practices Policies*

Stanley Pogrow[1]
San Francisco State University & University of Arizona
United States

Citation: Pogrow, S. (2017). The failure of the U.S. education research establishment to identify effective practices: Beware effective practices policies. *Education Policy Analysis Archives, 25*(5). http://dx.doi.org/10.14507/epaa.25.2517 This article is a substantially shortened version of the original, which went through two rounds of peer-review.

Abstract: One of the major successes of advanced quantitative methods has been its seeming ability to provide unbiased determinations of which education practices are effective for education in general and for improving the educational achievement and opportunity of the neediest students. The power of this methodology as applied in the top education research journals has led to periodic implementation of federal and state *effective practices policies*. In such policies the government or its proxy determines which programs are effective and then requires or encourages schools to spend its funds exclusively on those proven programs. For example, the federal Investing in Innovation (i3) initiative requires those applying for its largest dissemination grants to have had their intervention validated as effective by the What Works Clearinghouse (WWC). Some are now even advocating that all expenditures of federal funds by school districts be restricted to purchasing programs that research has proven to be effective, i.e., that they work. While this seems like rational policy on the

[1] *Conflict of Interest Statement*: Stanley Pogrow is the developer of the Higher Order Thinking Skills (HOTS) project, a specialized thinking development approach for Title I and Learning Disabled students in grades 4-8.

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 22/8/2015
Revisions received: 19/6/2016
Original acceptance: 21/6/2016; shortened version accepted 29/12/2016

surface, between 1998 and 2002 I produced a series of articles that showed that the most research validated program to develop the reading skills of students born into poverty was not actually effective in practice. Was this dichotomy between published research proving something to be effective and what was happening in the real world an anomaly or a more widespread problem? This article (a) analyzes in easy-to-understand language the validity of the gold standard scientific methodology used by the top research journals and WWC to determine whether practices are effective, and (b) examines the history of *effective practices policies* and their actual effectiveness. I conclude that the increasingly sophisticated methods used to assess the effectiveness of practices (a) are flawed and exaggerate actual effectiveness, and (b) do not provide the type of information practitioners need. As a result, research on effective practices tends to mislead rather than inform practice and are a major reason why efforts to reform high-poverty schools have had limited success. I therefore conclude that *effective practices policies* should not be implemented. I then suggest ideas for reforming the scientific process used to assess the effectiveness of education interventions.
Keywords: Equal educational opportunity; urban education; compensatory education; literacy; educational evaluation/assessment; school reform; evidence-based decision-making

El fracaso del establecimiento investigativo de educación de los Estados Unidos de identificar prácticas efectivas: Use precaución con las póliticas de prácticas efectivas

Resumen: Uno de los más grandes éxitos de los métodos cuantitativos avanzados han sido su aparente capacidad de proveer determinaciones imparciales de cuáles son las prácticas educativas eficaces para la educación y para mejorar el rendimiento académico y las oportunidades de los estudiantes más necesitados. Tal como se aplica en las revistas de investigación de la educación más destacadas, el poder ha resultado en la implementación federal y estatal de *póliticas de prácticas efectivas*. En tales políticas, el gobierno o su representante determinan qué programas son efectivos y después requieren que las escuelas gasten sus fondos exclusivamente en programas científicamente comprobados. Por ejemplo, una iniciativa federal requiere que aquellos que solicitan subvención del gobierno hayan tenido su intervención validada como efectiva por el *What Works Clearinghouse* (WWW). Al momento algunos están proponiendo que todos los fondos federales se restrinjan a los programas que la investigación ha demostrado ser efectivos. Mientras parece ser una pólitica racional, entre 1998 y 2002 produje una serie de artículos que demuestran que el programa más validado para desarrollar las habilidades de lectura de los estudiantes nacidos en pobreza no era realmente efectivo en práctica. ¿Sera que esta dicotomía que dice que las investigaciones publicadas y científicamente comprobadas que han demostrado ser efectivas son el estándar de oro en el mundo real, es una anomalía única o será una manifestación más perversiva? Este artículo (a) analiza en un lenguaje fácil de entender la validez de la metodología científica estandarizada y utilizada por las revistas principales de investigación y el WWC para determinar si las prácticas son efectivas, y (b) examina el historial de *póliticas de prácticas efectivas* y su eficacia real y actual. Concluyo que los métodos cuales son cada vez más sofisticados utilizados para evaluar la eficacia de las prácticas (a) son defectuosos y exageran la efectividad real, y (b) no proporcionan el tipo de información que necesitan los profesionales. Como resultado, la investigación sobre prácticas engaña en lugar de informar a la práctica y son una de las razones principales por las que los esfuerzos para reformar las escuelas de alta pobreza han tenido un éxito limitado. Por lo tanto, concluyo que las *póliticas de prácticas efectivas* no deben ser implementadas y sugiero ideas para reformar el proceso científico utilizado para evaluar la eficacia de las intervenciones educativas.
Palabras clave: Igualdad de oportunidades educativas; educación urbana; educación compensatoria; alfabetismo; evaluación; reforma escolar; toma de decisiones basadas en evidencia

O fracasso do estabelecimento dos Estados Unidos pesquisa em educação para identificar práticas eficazes: Tenha cuidado com as políticas práticas eficazes
Resumo: Um dos maiores sucessos de métodos quantitativos avançados têm sido sua aparente capacidade para fornecer determinações imparciais de que práticas educacionais eficazes para a educação e para melhorar o desempenho académico e oportunidades para os alunos mais necessitados. Quando aplicado em periódicos da educação mais notável de pesquisa, poder resultou na implementação de políticas federais e estaduais de práticas eficazes. Em essas políticas, o governo ou o seu representante determinar quais programas são eficazes e, em seguida, exigem que as escolas de gastar os seus fundos exclusivamente em programas cientificamente comprovados. Por exemplo, uma iniciativa federal exige aqueles que procuram subsídio do governo foram validados tão eficaz a sua intervenção pelo o *What Works Clearinghouse* (WWW). Quando alguns estão propondo que todos os fundos federais são restritas a programas que a investigação tem demonstrado ser eficaz. Embora pareça uma política racional entre 1998 e 2002, produziu uma série de artigos que mostram que a validado para desenvolver as habilidades de leitura dos estudantes nascidos no programa de pobreza não foi muito eficaz na prática. Será que esta dicotomia que diz pesquisa publicada e comprovado cientificamente que provaram ser eficazes são o padrão ouro no mundo real, é uma anomalia única ou ser um perversiva manifestação? Este artigo (a) analisa em uma linguagem fácil de entender a validade da metodologia científica padronizada e utilizado pelos principais revistas científicas e WWC para determinar se as práticas são eficazes, e (b) examina a história das políticas de práticas eficazes ea sua eficácia real e atual. Concluo que os métodos que são cada vez mais sofisticados utilizados para avaliar a eficácia das práticas (a) são defeituosos e exagerar a eficácia real, e (b) não fornecem o tipo de informação necessária por profissionais. Como resultado, a investigação sobre a prática engana ao invés de informar a prática e são uma das principais razões pelas quais os esforços para reformar as escolas de alta pobreza tiveram sucesso limitado. Portanto, concluo que as políticas eficazes não deve ser implementado práticas e sugerir ideias para a reforma do processo científico utilizado para avaliar a eficácia das intervenções educativas.
Palavras-chave: Oportunidades educacionais iguais; educação urbana; educação compensatória; alfabetização; avaliação; reforma da escola; a tomada de decisões baseada em evidências

Introduction

The research community has long argued that the failure to improve the quality of education for students born into poverty and to reduce academic inequities stems from practitioners not using the latest research on effective practices. As a result periodic policy initiatives require or encourage schools to use federal and state monies to adopt practices deemed to be effective by the research community. Over time such policies, hereafter referred to as *effective practices policies,*[2] have been enacted and the criteria for certifying that a program or practice is effective have become more rigorous and institutionalized within government agencies. Historically, such policies have ranged from merely publishing lists of effective programs for practitioners and providing funding for the dissemination of such programs to mandating that some state and/or federal funds be only used for programs certified as effective. There is now a push to further strengthen existing *effective practices policies* by requiring that *all* federal funds be spent only on scientifically certified effective practices.

[2] The terms practices, interventions, and programs will be used interchangeably, and the term *effective practices policies* will refer to any formal approach designed to improve student and school performance regardless of whether it is a highly detailed program such as *Accelerated Schools*, *Cognitive Tutor*, etc., or a more general practice or intervention, such as providing positive reinforcement to struggling students, merit pay, etc.

On the surface, policies promoting the use of research-validated practices would seem to be a major success for the education research community and for the professionalization of education practice—as well as an efficient, commonsense way to improve education. However, this article shows how rather than ushering in an era of improved practice and student achievement, the fiscal incentives and high stakes of getting certified as an effective practice ushered in opportunism that had the opposite effect.

The Emergence of *Effective Practices Policies*

The driving force behind *effective practices policies* was the growing dissatisfaction with the effectiveness of federal education programs; particularly Title I of ESEA, the largest federal education program. Title I provides supplemental funds for school districts with significant pockets of poverty, offering extra assistance to children born into poverty and thereby reducing achievement gaps.[3] Since the program was initiated in 1965, periodic evaluations of Title I have shown little effect on academic achievement. Frustration with the lack of progress, combined with the push for a more business-like approach to education, propelled the idea that Title I schools should use scientifically validated programs to achieve the desired improvements.

Establishing a Scientific Methodology to Determine Program Effectiveness

Implementing *effective practices policies* requires a specified methodology for determining that a program or practice is effective. The earliest federal efforts to establish a government seal of approval that certified programs as effective was the National Diffusion Network (NDN), which existed from 1974 through 1995. NDN also provided funds to disseminate the programs it judged to be effective. Program evaluation centers were also established in regional labs funded by the U.S. Department of Education to inform districts about effective programs

Criticism began to mount that these early systems for certifying and disseminating effective practices were too lax and did not use rigorous standards of evidence, and that too many interventions were being designated as effective. In order to better advise practitioners as to which interventions are effective based on exemplary research, Congress established the What Works Clearinghouse within the U.S. Department of Education's Institute of Education Sciences in 2002. The function of this institute is to provide "rigorous and relevant evidence on what works, what doesn't, and why, to improve educational outcomes for all students, particularly those at risk of failure."[4] The role of the What Works Clearinghouse is to set rigorous quantitative standards based on the best science to provide guidance for practitioners as to what works by assessing the quality of research evidence supporting a given intervention. The What Works Clearinghouse also issues a rating on whether the evidence supporting an intervention does or does not meets its standard of evidence. (There is an intermediary rating that the evidence meets its standard with reservations.) Those interventions that meet its standards are presumed to work; i.e., to be effective. The Clearinghouse was conceived to perform the same scientifically rigorous review function in education as the Food and Drug Administration (FDA) does to validate scientific evidence of the

[3] Title I has grown to approximately $14.4 billion for the 2015 school year and is expected to provide supplemental help to 23 million students.

[4] The mission statement for the Institute of Education Sciences can be found at: https://ies.ed.gov/aboutus/. The mission statement for the What Works Clearinghouse can be found at: https://ies.ed.gov/ncee/wwc/FWW, and the criteria for reviewing studies (Version 3.0) is at: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

effectiveness of drugs. The Institute of Education Sciences began to provide funds for program developers to conduct rigorous research on the effectiveness of their interventions.

As a result of the desire to have a more rigorous method to determine which programs were effective, the top research journals and funding agencies began to require more sophisticated research designs and statistical analyses in articles and proposal submissions. Statistical standards became increasingly complex as a requirement for (a) getting quantitative research published in the top journals, (b) obtaining government funding to test an intervention, and (c) getting the evidence to prove that a program is effective and obtain certification by the What Works Clearinghouse.

Evolution of Effective Practices Policies

Effective practices policies range from simply providing evidence as to what programs are effective to mandating that schools can only use those programs that have been deemed to be effective and restricting the use of specific funds to such programs. In between those extremes are policies that provide monies to developers to disseminate programs deemed to be effective or that provide extra monies as an incentive for schools to adopt an effective program.

The earlier efforts to promote effective practices, such as the National Diffusion Network, were limited because school participation was voluntary and only tiny amounts of funding were available to help developers disseminate their programs. The first federal *effective practices policy* that restricted the use of funds by schools occurred in 1997 when Congress passed the Comprehensive School Reform Demonstration (CSRD) law, also known as Obey-Porter. This law allocated $150 million dollars for schools to adopt research-based reform programs. The guidelines for this law contained a list of "effective" comprehensive programs that met the goals of the type of reform that the legislation required.[5] To my knowledge this is the first time that federal monies were restricted in this fashion. Datnow (2000) reported that by 2000 over 1800 schools had each received $50,000 a year for up to three years for adopting one of the listed effective programs. There was also an effort in a number of states to require schools to use their Title I funds to adopt one of these "effective" programs. During this same period the New American Schools (NAS) project was established, raising private monies for the development and use of one of seven comprehensive school reform models.[6]

The most powerfully restrictive example of an *effective practices policy* occurred in New Jersey. The state of New Jersey had extremely large differences in funding between rich and poor school districts. After more than 20 years of lawsuits claiming that such disparities violated its state's constitution, the New Jersey Supreme Court finally agreed. It decreed in *Abbott v. Burke* (1997) that the state had to invest approximately a quarter of a billion dollars in 1997-98 to bring the spending of the 440 poorest schools in the state up to the levels of the richest. This was an unprecedented and bold effort to reform public education. However, one of the Supreme Court judges went a step further to insure that the new state monies were spent effectively. As a result, New Jersey initially planned to require that all Abbott elementary schools use this substantial amount of new monies on a single program, which it considered to be the most effective. The state then allowed the Abbott schools to also select from a few other "proven" programs if they could explain why the "best" one did not meet their needs.

---

[5] The federal legislation did not *require* that all schools use one of the programs on the list, but the existence of the list meant that the programs on it were presumed to be the best by schools and states.

[6] All of the NAS models were on the Obey-Porter list even though most were new and had no established track record. These newer models included: Audrey Cohen, ATLAS, Co-NECT, Expeditionary Learning, Modern Red Schoolhouse, and National Alliance for Restructuring Education (America's Choice).

With the What Works Clearinghouse, the Institute of Education Sciences, and the use of more rigorous statistical research methods in place, in 2009 the Obama administration decided to promote, but not require, the use of research validated programs. It established the Investing in Innovation (i3) fund in the U.S. Department of Education's Office of Innovation and Improvement. The purpose of i3 was to fund innovative initiatives with a record of improving student achievement and attainment, as determined by What Works Clearinghouse, to scale up their adoption. This leveraged the importance of What Works Clearinghouse's seal of approval and more monies were provided to those interventions that had the strongest evidence of effectiveness. In 2010 the highest level of dissemination grants was $50 million per program. The stakes had been raised to get What Works Clearinghouse's highest rating, and a form of "ranking regime" (Gonzales & Núñez, 2014) had been institutionalized.

Currently, the certification and utilization of effective practices applies only to developers and researchers seeking government funding and/or validation. Time will tell if, as some advocate, schools will be required to use state and/or federal funds only for interventions whose evidence of effectiveness have been approved by the What Works Clearinghouse.

On the surface, this is a story of the apparent triumph of rational science and policymaking and a clear validation of the importance and applicability of educational research for improving practice. However, there are major problems with the rigorous scientific methodology used to determine the effectiveness of interventions.

Methodological Problems

Reliance on Relative Comparisons

Research on program effectiveness uses the awesome power of modern statistical analysis to determine whether experimental students are doing better than those in comparison schools. However, there is a problem with basing evidence *solely* on relative data from comparisons between experimental and comparison groups. A non-technical way to understand the problem of relying on relative comparisons is to consider the following vignette of a couple living in the upper Midwest during a particularly bad cold spell deciding where to vacation in January. The goal is to find a place where they can relax on a warm beach and get a tan:

Wife: *I cannot wait for our vacation in January. Let's go somewhere warm.*
Husband: *Definitely.*
Wife: *Where should we go?*
Husband: *I just read that Greenland is warmer than Antarctica in January.*
Wife: *That sounds great.*
Husband: *Even better, due to climate warming Greenland will be warmer this year than last. Plus, it has 27,394 miles of coastline, so it will be no problem finding beaches.*
Wife*: That's great. It will be wonderful to go somewhere where we can leave our winter clothes behind.*

Their decision is certainly evidence-based—but it is clearly a lousy decision. Why was this couple's *evidence-based decision* so bad? It was bad because they relied *solely* on relative data. They needed a key piece of absolute data, which in this case was the actual temperature in Greenland in January. The actual temperature is -8 C with zero hours of sunshine. This couple is far more likely to die from hypothermia in Greenland in January than to get a tan. With the right actual/absolute data the couple would arrive at the correct decision to reject both of these choices as it would be warmer to stay at home or to seek other options.

In other words, you cannot make intelligent decisions relying *solely* on relative data (i.e., data that relates external events/situations to each other) no matter how compelling that evidence appears to be. You will always need some absolute data. The key piece of absolute data needed to judge the quality of a program is the answer to the obvious question: *How did the students in the experimental intervention actually perform?* For a program to be judged effective we would expect students to do reasonably well on an absolute basis. Reasonable people can debate as to what an expectation of students "doing well" is. However, increasingly published research does not report how the experimental students actually performed. Under current conceptions of rigorous science, top research journals and the What Works Clearinghouse rely *solely* on the relative difference between an experimental and comparison group to determine the effectiveness of an intervention. The amount of relative difference is expressed as an *effect size* (ES).

Relying on Adjusted Outcome Scores

Compounding the problem of relying on relative scores is that the reported outcome scores are usually not actual outcome scores but *adjusted* ones.

Post-test scores are typically adjusted based on initial differences between the groups. So for example, if the initial reading scores of the experimental group are lower than the scores of the comparison group, then the post-test scores of the experimental group are statistically adjusted relatively upwards using the *analysis of covariance* statistical procedure. This is statistically legal. However, if school leaders adjust their students' scores, chances are that they will be fired and possibly go to jail. So when researchers report scores that have been statistically adjusted relatively upwards it is common and helpful to practitioners for the researchers to also report the unadjusted scores. However, in the research I studied only adjusted outcomes were reported, and the adjustments increased the results for the experimental groups.

The other way that scores are adjusted is to transform everything into normalized Z scores. Without going into a statistical explanation, this transformation enables one to compare apples and oranges. This is actually very useful as it enables a researcher to compare scores from different state tests or from different units, such as comparing the relative impact of height, income, and weight on some outcome. So the relative results often reflect comparisons of the means of adjusted scores, or adjusted normalized scores. Sometimes additional adjustments are made.

It is not unusual to find articles published in AERA journals such as Pane, Griffin, McCaffrey, and Karam (2014) examining the effectiveness of *Cognitive Tutor for Algebra I* that have extensive tables of actual mean scores of the experimental and control groups. However, these are all pretest scores. They are included primarily for the express purposes of determining whether adjustments are warranted. However, once all the relative comparisons are reported, the only post-test scores shown in this study and others are the adjusted ones. So there is no indication as to what the actual performance of the experimental students were.

Overstating the Practical Importance of Relative Differences

Given the absence of actual results on how students are actually performing in published research, researchers and the WWC rely on the relative difference between the performance of students or schools using the intervention as compared to those not using it to determine whether the intervention is effective. As already noted, it is problematic for either the vacationing couple or practitioners to make decisions based solely on relative data. However, the situation is even worse given the amount of relative difference that researchers use to conclude that an intervention is effective.

Originally, researchers determined a program's effectiveness by using the criterion of whether a difference favoring the experimental group was statistically significant (e.g., $p<.05$). However, that criterion merely indicated whether the relative difference could have occurred by chance. The problem was that this criterion did not indicate whether the difference was substantial, or beneficial, enough to be of practical importance.[7] As a result, the preferred criterion to characterize the importance of the relative differences favoring the experimental group came to be effect size (ES), sometimes also referred to as Standard Deviation Units (SDU).

Researchers generally consider an effect size (ES) of .2 as the point at which the relative difference between groups becomes large enough to have practical significance; that is, the experimental program can be judged to be effective. (Don't worry about how this value is calculated.) Using .2 as the minimum cutoff for indicating that a relative ES difference is important is based on the work of Cohen (1988). However, researchers forget that Cohen characterized such a small difference as "difficult to detect." In the real world it makes no sense to expect leaders to recommend that their school(s) adopt an expensive, time-consuming program in order to produce improvements that are at best "difficult to detect." (It is only at .5—or more than twice as much—that Cohen characterizes differences as becoming "visible to the naked eye"; i.e., apparent to practitioners and community members.) Indeed, at .2, there is substantial overlap between the performance of the experimental and comparison groups, and only a very small percentage of students are benefitting.[8]

The following discussion of effect size results simply relates them to the benchmark of .2—so that an effect size of .1 means that the experimental program produced a benefit that was half of "difficult to detect;" i.e., almost impossible to detect. (However, keep in mind that regardless of how big the ES is it does not describe actual performance—only relative performance.) Unfortunately,

---

[7] See Carver (1978) for a discussion of the limitations of relying on statistical significance to indicate whether a meaningful difference has occurred between groups.

[8] Small effect sizes can have theoretical importance. However, the question here is whether they have practical importance in terms of identifying effective interventions.

There are two main arguments as to why .2 should be considered an important effect size (ES) for education practice. Lipsey et al. (2012) correctly noted that Cohen's definition was across all the social sciences. They argued that each field should have its own threshold based on what the research experience has been in that field. Since the effect sizes of comprehensive interventions in education have come in below .2, education should use a lower threshold as to what the minimum effect size is for considering a program to be effective. The problem with this rationale is that it is a relative one. The fact that some intervention is shown to be a bit better than other programs of dubious effectiveness does not mean that it is actually effective. In the end a small difference is a small difference and simply because educational researchers have not figured out to date how to produce larger effect sizes simply means that better interventions need to be developed.

The second argument for using lower ES cutoff points is that the medical community approves the use of drugs, such as baby aspirin to prevent fatal heart attacks, despite tiny ESs. Aside from the obvious problem that the functioning of the human body differs from the functioning of complex social organizations such as schools, there are also many other differences between using baby aspirin as compared to implementing a complex educational intervention. For example, baby aspirin costs pennies a day and does not require training or a learning curve. In addition, Carroll & Frankt (2015) note that if 2,000 people who have a 10% chance of a heart attack take aspirin for two years only one heart attack is prevented, and the other 1,999 are unaffected. Would practitioners agree that an expensive, time-consuming intervention that has no benefits for 99% of their at-risk students should be implemented?

researchers and reporters tend to overstate the importance of research with low ES's. Consider the following example.

The headline of the results a study of the effects of an i3 funded reading program used in Title I schools in *Education Week* declared: "School Improvement Model Shows Promise in First i3 Evaluation" (Sparks, 2013, p. 8). This headline parroted the conclusion of the researchers. The conclusion of "promise" is based on the following results in kindergarten shown in Table 1.

Table 1
Effect Sizes for Kindergarten Reading Outcomes

Outcome	ES
Letter-Word Identification	-0.01
Word Attack	0.18

Why is such a result "especially promising?" Given that the earliest grades are the easiest ones to produce big effect sizes it is surprising to see a negative effect size, which means that the experimental students did a tiny bit worse. The one positive result is slightly less than "difficult to detect." The positive evaluation is based on the fact that the .18 for a single sub-skill is larger than Title I's overall effect size of .11.

However, not only do such statements by researchers and reporters overstate the importance of these effect sizes, they also indicate a lack of knowledge about the state-of-the-art of research on early elementary reading interventions—which have found ES's that are two to three times as large. Hattie (2009) reviewed all 800+ existing meta-analyses on education outcomes. These meta-analyses include approximately 52,000 studies. His finding was that the average effect size for all educational interventions studied was .4. He then argued that this should be considered as a cutoff for being considered as producing "real world" effects, particularly for expensive programs. More specifically, Table 2 shows the ESs for the types of reading outcomes most similar to those reported for the i3 funded program. These effect sizes are consistently positive and three times the size of those being praised by the reporter.

Table 2
Hattie's Effect Sizes for Reading Outcomes

Intervention Outcomes	ES
Vocabulary Programs	.67
Phonics Instruction	.60
Repeated Reading	.67

However, to be fair, there is no breakdown of Hattie's effect size calculations as to (a) grade level, (b) how intensive the interventions were, or (c) whether the results were measured on a standardized/end of year test (which tend to produce lower effect sizes). As a result, the best

comparison for the effect sizes reported in Table 1 is the meta-analysis of the effects of intensive reading interventions (at least 100 days duration) for K-3 *struggling* readers by Wanzek et al. (2013) and Scammacca et al. (2007). Table 3 below lists the meta-analysis results for those reading skills most closely related to those reported for the i3 funded program:

Table 3
Effect Sizes for Meta-Analyses of Intensive Reading Interventions for Grades K-3

Reading Outcome	*N*	ES
Comprehension outcomes	25	.46
Reading fluency	11	.34
Word reading	53	.56
Word fluency	18	.56
Spelling	24	.40

All of these ESs are consistently positive and most represent a "clearly visible" relative advantage for the interventions being studied. The characterization of the results for the i3 funded reading program is simply wrong. An effect size of .18, even if unadjusted, provides no rationale for schools to adopt the program or for the government to help disseminate it. This is equivalent to expecting people to get rid of their existing car and buy a new one simply because it gets one more mile per gallon. The difference has to be larger to affect personal behavior—and the effect size for an expensive program in reading at the elementary level has to be substantially larger than .2 for leaders to be encouraged to adopt it, or for it to be considered effective, i.e., proven to work.

In addition, researchers are pressing to further lower the minimum ES threshold for considering an intervention to be effective to below .2. For example, Borman, Grigg, and Hanselman (2016) concluded that self-affirmation exercises are a good way to reduce the achievement gap in mathematics because they obtained effect sizes of .09 and .11 (i.e., half of "difficult to detect"), and these are consistent with the effect sizes of other national reforms. So because a new trivial effect size is the same as other so-called "successful" reforms, this is taken as an indication that this new intervention is also effective. The correct interpretation of such relative comparisons, as with the example of our hapless couple seeking a warm vacation spot, is that probably none of these reforms are desirable choices.

Using Invalid Hypothetical Extrapolations to Explain Unintuitive Results

Another problem with interpreting relative effect sizes is that they often are comparing adjusted normalized numbers. However, these are just abstract numbers that have no recognizable real-world meaning to practitioners or policy makers. So while a relative effect size can calculate the amount of difference between these abstract numbers, what does the result tell you about what happened to students or schools? It is therefore common practice in the literature to transform the effect size results into something familiar to practitioners and policy makers; hereafter referred to as *extrapolations*. The most common type of extrapolation is to equate the effect size to changes in test scores with statements such as: "The size of the effect favoring the experimental students is the equivalent of moving students from the 50th to the 58th percentile on the Stanford Reading Test." However, this seemingly impressive improvement is only a hypothetical relative extrapolation

because we do not know whether students actually scored at the 58th percentile or the 18th. In these types of *hypothetical extrapolations* in research journals students may not have even actually taken the test mentioned or any nationally normed test. In addition, this type of extrapolation assumes that the distribution of scores for students at risk is the same as national norms, which is clearly not true.

Another type of hypothetical extrapolation converts effect sizes into extra days of learning. For example, the widely cited result from the CREDO (2013) study at Stanford University combined scores from 16 state tests to test whether students did better at charter or traditional public schools (TPS). This study found that black and Hispanic students made 14 days of additional learning per year in charter schools as compared to TPS.[9] This is an impressive difference that suggests that charters are superior. However, the effect size of the difference produced by these two types of schools was only .02, or one-tenth of "difficult to detect." Common sense would suggest a "difficult to detect" difference would be about three extra days of learning and one-tenth of that would be roughly one-third of a day; or about two hours. While most statisticians would scoff at this simplistic heuristic, the question remains: Which is the correct extrapolation; about two hours, which is not a substantial difference; or 14 days, which is substantial? It turns out that two hours is closer to the truth. First, CREDO (2013) noted that their findings "…are only an estimate and should be used as a general guide rather than as empirical transformations" (p. 13). CREDO is essentially admitting that there is no real empirical basis for their published extrapolation. Maul and McClelland (2013) noted that CREDO's conversion of effect size into days of learning was "insufficiently justified" and that there really was not a substantial difference between the performances of the two types of schools. Second, another source confirms that CREDO's extrapolation was a gross over-estimate. In Gene Glass's evaluation research an effect size of .02 equates to only 2.4 days of learning (G. Glass, personal correspondence, June 20, 2016). However, this is for a nationally normed sample. Since this CREDO sample was below national norms I will arbitrarily subtract a day from Glass's estimate, and be left with a difference of 1.4 days of learning. The result is now relatively close to what was generated from the "difficult to detect" heuristic (as compared to the CREDO result).

In other words, CREDO's published results were based on a "misleading extrapolation" and led to the wrong policy conclusion; one that denigrated traditional public schools. The fact is that all of these extrapolations are really only heuristics. Glass also viewed his calculation as a heuristic (G. Glass, personal communication, June 20, 2016). In other words, no one knows precisely what effect sizes generated from adjusted normalized relative data equate to in terms of actual days of learning or test scores for non-normed samples. This is especially true for non-normed measures. In this vacuum Cohen's heuristic of "difficult to detect" provides a useful and easy way for practitioners and policy analysts to accurately spot when research uses a "misleading extrapolation" to exaggerate the practical importance of its findings.[10]

Hypothetical extrapolations are no substitute for knowing how the experimental students actually performed. Such extrapolations are akin to telling our hapless vacationers that the difference between the temperatures in Antarctica and Greenland in January is equivalent to raising the temperature in Miami in January from 76 to 85 degrees. This hypothetical extrapolation makes Greenland seem warm! At the end of the day the reality is that the temperature in Greenland in January is not 80 degrees—it is still -8.

---

[9] These are the results from schools that continued from a prior analysis as opposed to the new charter schools in this study.

[10] This method also suggests that the belief that an ES of .1 equates to a month of extra learning in a year favoring the experimental group is similarly a "misleading extrapolation", as an extra month would probably be a noticeable difference.

A Summation of How Research Distorts the Effectiveness of Interventions

The examples cited above show how the current research methodology in top journals distorts and exaggerates the actual effectiveness of interventions. The current predominant statistical approach for determining program effectiveness relies on external relative comparisons of sets of highly adjusted data using statistical criteria that extols the importance of differences that are difficult to detect in actual practice, and that then require hypothetical extrapolations of questionable validity to be converted into results that practitioners and policy makers can understand. This convoluted process makes it easy for a statistically savvy researcher to make an ineffective program appear to be successful via strategic adjustments—either intentionally or unintentionally. It is a house of cards built on the basis that a given trivial ES is bigger than some other trivial effect size. This methodology leads to confusion in the research and journalistic communities as to whether programs are producing *actual* (i.e., unadjusted, non-relative, non-normalized) improvement levels of student performance that are apparent in the real world. These methodological problems render current efforts to use scientific evidence to determine what works highly problematic. These problems provide a basis for explaining why prior iterations of implementing *effective practices policies* had little effect. However, that has not stopped the ongoing push for such policies.

The Continued Promotion of Effective Practices Policies

There is no evidence that *effective practices policies* work. Good, Burross, and McCaslin (2005) studied the effects of the Obey-Porter *effective practices policy*. Their study concluded that the mandated use of highly effective comprehensive school reform models (CSR) did not provide any advantage as compared to what practitioners were doing without the extra funding. This study's findings are similar to the conclusion of the later Burdunny et al. (2009) study that found no benefit from expert-selected programs over what practitioners were already doing. In other words, such policies only benefit the vendors whose programs had been "proven to be effective."

However, this has not stopped highly respected researchers and policy makers from advocating for policies requiring the use of practices that have been proven to be effective as essential for producing gains on the national level. Such advocacy has appeared in highly prestigious and visible platforms. The most common advocacy for a federal role is that Title I, the $14 billion program to help high-poverty schools, should either promote or require that these federal funds be used for programs that have been proven to work. Some have gone even further. For example, in an *Education Week* commentary Cross et al. (2014) advocated:

> … the impact of the federal investment in education R&D could be significantly improved if, among other things, a reauthorized ESEA elevates the Institute of Education Sciences to a lead position in the evaluation of *all* federal education programs…and promotes the use of *proven programs in all department grant funding.* [Emphasis added]

This raises the bar by advocating that the federal government require the use of "proven practices" for *all* programs—not just for Title I. However, given (a) the lack of evidence that *effective practices policies* actually improve student outcomes, and (b) the disconnect between actual real-world effectiveness and how researchers determine that a practice is effective, implementing any form of *effective practices policies* at this point in time is an unwarranted governmental intrusion into local educational decision-making that will result in stagnation and deflect efforts to seek alternative approaches to developing and identifying practices that are actually effective.

Discussion

Is the Current "Rigorous" Scientific Approach for Identifying Effective Practices Useful or Valid?

This article has highlighted a series of general methodological problems and artificialities in how we use research to identify effective practices; e.g., over-reliance on *relative* versus *actual* performance, hypothetical extrapolations, and effect sizes that are "difficult to detect." My concern about the validity of the methodology used to prove that a given practice is effective started with my earlier research (Pogrow, 1998, 1999, 2000a 2000b, 2002) that demonstrated that a widely used research validated program was not actually being successful in the schools and districts that I studied—including some of those where the program had been found to be successful in published research. This led me to wonder whether this dichotomy between published research and what was actually happening in these schools and districts was an anomaly or reflected a widespread problem with how we assess the effectiveness of programs in education. In other words, can we trust any of the findings of the What Works Clearinghouse about which programs work or any other list of "research validated" or "proven to work" practices?

There is a growing chorus of researchers criticizing the emphasis on rigorous, "gold standard" Randomized Controlled Trial (RCT) research methodology to inform practice. Ginsburg and Smith (2016) examined the evidence for all the math programs certified by the What Works Clearinghouse as having evidence of effectiveness based on RCT research. They reviewed all 18 math programs that had been certified by the Clearinghouse, which comprised a total of 27 approved RCT studies. They found 12 potential threats to the usefulness of these studies and concluded "…none of the RCT's provides useful information for consumers wishing to make informed judgments about what mathematics curriculum to purchase" (p. 44). Some of the threats overlapped those discussed in this article, such as greater amounts of instructional time provided to help the experimental group.

Bryk, Gomez, Grunow, and LeMahieu (2015) have also critiqued the usefulness of RCT approaches in education. Gopal and Schorr (2016) argue that RCT analysis is unlikely to be helpful because the types of interventions that are easily studied within clear controlled experiments are probably too simplistic to solve complex problems.

In addition to questions about the usefulness and validity of findings from gold-standard methodology, there is also growing recognition of problems with relying on small effect sizes. Where Ginsburg and Smith (2016) were able to determine the error effects of a threat, they found that the error generated by even one of those threats was at least as great as the effect size favoring the treatment group. This is even more problematic for the instances where there were multiple threats. It is also likely that the types of adjustments to data and hypothetical extrapolations add even more error. All of these taken together make the actual effect indeed "difficult to detect." In addition, the problems of basing education practice on small effect sizes also extend to the results of meta-analyses. Glass (2016) found that in contrast to meta-analyses of medical trials wherein the effect sizes of the individual studies are fairly similar, in education the variation of effect sizes within a given meta-analysis is "great" and the effects are "relatively small" (p. 71). As a result, Glass concluded that meta-analysis in education has failed to provide clear policy guidance.

Indeed, relying on small effect sizes appears to be a problem even in the field of psychology from which many of the current techniques for assessing the effectiveness of education interventions are drawn. Psychology was recently rocked by the finding that *more than 60% of the key research findings* that form the basis of many of its practices *could not be replicated* (Carey, 2015*)*. In this case fraud was not the issue. Open Science Collaboration (2015) found that "…reproducibility

success was correlated with the strength of initial evidence …such as larger effect sizes" (p. aac4716-6). If findings of psychology research with small effect sizes cannot be replicated in the lab, how can we expect the research recommendations of the What Works Clearinghouse, and applied education research in general, that use the same techniques and statistical criteria to be replicable in schools? If the findings of a program are not replicable in schools it should not be considered "effective." Ioannodis (2005) went a step further. This Professor of Medicine and of Health Research and Policy at Stanford University concluded that the smaller the effect sizes in any scientific research the less likely it is that the research findings are true.

Similar problems have been found in psychiatry. Kraemer (2016) noted that when RCT trials of experimental treatments are conducted at various sites the results often do not replicate, and that when clinicians apply the results from such rigorous studies the results they obtain do not match what researchers report. If the results from randomized trials do not reflect real world outcomes in the relatively simple relationship between a clinician and an individual patient, how is such methodology going to reflect actual outcomes in the far more complicated network of social interactions that exists in schools?

So it is starting to appear that the existing scientific conventions in applied education research for identifying effective practices used by top research journals and government agencies are flawed and lead to misleading conclusions across disciplines—even when the research has high levels of integrity. It is simply impossible, as the Ginsburg and Smith (2016) research suggested, to control for all the confounding variables in the chaotic world of educational practice regardless of how much money or time is spent. This means that you can never be sure that the program being studied caused the reported relative differences, and whether those differences are real or have practical significance for practitioners. The current complex statistical methods generate a hypothetical relativistic mathematical system—as opposed to findings of clearly visible improvements in the real world that practitioners seek. This is especially true when in the end we do not know how students actually performed in the studies and rely just on relative external comparisons. So in the end practitioners and policy-makers are left in the same situation as the hapless couple trying to plan a vacation. Practitioners end up not knowing whether either of the choices in the studies produced better results than currently exist in their school(s), and policymakers end up not knowing whether the experimental group with the "effective" treatment actually performed terribly.

Taken together it is highly questionable whether the method of research that has monopolized research efforts at all levels is useful or valid for informing the decisions of practitioners. As a result, there is currently no basis for any form of *effective practices policies*. Given the evidence to date, *effective practices policies* are far more likely to harm education than improve it. For example, the single program that New Jersey pushed all Abbott elementary schools to use was the one my research found to be actually ineffective. Therefore such *effective practice policies* should not be enacted. Rather, we need to go back to the drawing board on how to identify effective practices.

Alternative Scientific Methods for Identifying Effective Practices

Fortunately, the widespread belief that "rigorous science" can only be conducted with sophisticated experimental designs and RCT is wrong. A small group of educational researchers has begun to explore other forms of scientific methods for discovering, validating, and disseminating effective practices—methods that have produced major improvements in clinical practice. For example, Gawande (2007) noted that obstetrics, which has saved more lives than any other branch of medicine by identifying effective practices to reduce infant mortality, does not conduct experiments. *Improvement Science*, which has been successful in improving the delivery of health

services, does not use controlled trials to identify and disseminate effective practices (Berwick, 2008; Plsek, 1999). The latter includes the use of effective practices in hospitals, which like schools, are settings with complex social interactions. These newer scientific methods focus on actual improvement outcomes and use much simpler and more intuitive statistics.

Pogrow (2015) described how alternative approaches to scientific discovery have historically provided major spurts in knowledge. Some researchers have started to apply these alternative methodologies for developing and identifying effective approaches in education. Bryk, Gomez, Grunow, and LeMahieu (2015) and Pogrow (2015) have demonstrated major progress in improving a major problem of practice using an alternative scientific method by demonstrating substantial improvement at scale—without using experimental designs. These techniques seem more relevant to informing practice in a profession such as education with its complex social interactions and time-constrained improvement needs. Pogrow (2015) also recommended much more stringent analysis of actual performance results for determining whether a program is effective. Under these criteria *many/most* programs considered to meet existing criteria of best research evidence would not be considered to be effective—nor would most of the programs certified by the What Works Clearinghouse. Given all the problems that have been identified in the existing paradigm of applied research, and available alternatives, it is time to reconsider whether this conception of rigorous science that has been imposed on education is the best way to identify and develop effective programs. This means considering whether simpler research designs, even ones without control groups, using simpler measures of actual performance of experimental groups, are better able to predict the effectiveness of an intervention at scale.

Conclusions

The only thing worse than practitioners ignoring research that has truly demonstrated practices to be effective is for the research community to certify practices as being effective that are not. It is even worse when the research community encourages government to disseminate, or encourage/require practitioners to use, such practices. Alas, the methodology prized by the top research journals and government panels for identifying effective practices makes assumptions and adjustments that introduce artificialities and errors into the analysis. As a result, it is not clear that the studies produce useful or valid results about the effectiveness of practices. The methodology values findings of benefits that are "difficult to detect" in a relative hypothetical mathematical system—as opposed to findings of clearly visible improvements in the real world. As a result, there is now reason to question whether practitioners can trust any of the What Works Clearinghouse's recommendations; a conclusion that is supported by the findings of Ginsburg and Smith (2016) with respect to a wide range of programs. The same concern can probably be ascribed to other sources that provide "evidence" of what works and lists of proven practices.

As a result, it appears that we need to start over and rethink the methodological approach used to identify effective practices in research journals and government panels. We would probably do better to look at the simpler methods employed by obstetrics and the *improvement science* of health services. Both of these have established track records of identifying and disseminating effective practices that seem to have produced greater improvements in clinical practice than the experimental RCT model that education has viewed as the only model of rigorous science. Indeed, my research showed that the simpler methodologies used by school district research offices to determine actual program effectiveness were more valid than the results published in the top research journals for the same districts. In addition, any future methodology for certifying programs to be effective should put a premium on the actual, unadjusted performance of students receiving the treatment,

particularly at-risk subgroups, and whether the improvements (a) are ones that would likely be clearly visible to practitioners and parents, and (b) whether such improvements have occurred "reasonably" consistently in case studies. Such an approach would require a major culture change in the education research community. It would also require substantial changes in the design of graduate quantitative methods courses for practitioner preparation programs. Above all it would require practitioners, policymakers, and researchers to consider the likelihood that practices previously considered to have the most proven track records of scientific evidence may in fact not be effective.

In the meantime, government should suspend supporting the production or dissemination of any list of effective/proven programs, and the profession should resist the seductive calls for implementing any form of *effective practices policies*. Rather, we need to build policy around the principle that students born into poverty deserve the best teachers and the most rigorous curricula accompanied by the most creative and reflective forms of teaching and learning to inspire them to achieve to their full ability.

Acknowledgements

This work would not have been possible without the help of many. I am indebted to the following scholars (listed in alphabetical order) for the ideas that they contributed to this work: David Berliner, Christopher Bettinger, Arnold Danzig, Alan Ginsberg, Gene Glass, Michael Kirst, Marshall Smith, and Timothy Weekes. Thanks also to those at EPAA/AAPE, especially Audrey Amrein-Beardsley and Gustavo Fischman for their timely suggestions and encouragement, and to Stephanie McBride-Schreiner for polishing the final manuscript. The thorough and tough critiques of anonymous reviewers were essential to making the ideas deeper and better. Thanks also to Richard Venezky and Herbert Walberg for initiating this line of research, and to the many district researchers and teachers who shared their data and experiences with me. Finally, the hordes of typos expunged by Bruce Smith, Yisroel Shaw, Warren Ockrassa, and Deborah Sherman strengthened this work. Any remaining problems with the ideas or the writing are strictly my responsibility.

References

Abbott v. Burke, 149 N.J. 145, 693 A.2d 417 (1997) (Abbott IV).

Berwick D. M. (2008). The science of improvement. *JAMA, 299*(10), 1182-1184. https://doi.org/10.1001/jama.299.10.1182

Borman, G. D., Grigg, J., & Hanselman, P. (2016). An effort to close achievement gaps at scale through self-affirmation. *Educational Evaluation and Policy Analysis*, *38*(1), 21–42. https://doi.org/10.3102/0162373715581709

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to Improve: How America's Schools Can Get Better at Getting Better*. Cambridge, MA: Harvard Education Press.

Burdumy, J. S., Mansfield, W., Deke, J., Carey, N., Lugo-Gil, J., Hershey, A., . . . Pendleton, A. (2009, June 8). *Effectiveness of selected supplemental reading comprehension interventions: impacts on a first cohort of fifth-grade students*. Mathematica Inc., Presentation at *The Institute of Education Sciences* research conference. Prepared by *Mathematica Policy Research* as the lead contractor. Prepared for the National Center for Education Evaluation and Regional Assistance, The Institute of Education Sciences. U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20094032/pdf/20094032.pdf

Carey, B. (2015, August 27). Many psychology findings not as strong as claimed, study says. *New York Times*, retrieved from http://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html

Carroll, A. E., & Frakt, A. (2015, February 2). How to Measure a Medical Treatment's Potential for Harm, *New York Times*. Retrieved from http://www.nytimes.com/2015/02/03/upshot/how-to-measure-a-medical-treatments-potential-for-harm.html

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378-399. https://doi.org/10.17763/haer.48.3.t490261645281841

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

CREDO. (2013). National charter school study. Center for Research on Education Outcomes. *Stanford University,* Stanford, CA. Retrieved from http://credo.stanford.edu/documents/NCSS%202013%20Final%20Draft.pdf

Cross, T., Tamayo Jr., J. R., McKibben, S. & Goldstein, M. (2014, May 1). Making the most of federal ed. research. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2014/05/01/30cross.h33.html?qs=making+the+most+of+federal

Datnow, A. (2000). Power and politics in the adoption of school reform models. *Educational Evaluation and Policy Analysis*, *22*(4), 357-374. https://doi.org/10.3102/01623737022004357

Gawande, A. (2007). *Better: A surgeon's notes on performance*. New York, NY: Metropolitan Books.

Ginsburg, A., & Smith, M. S., (2016). Do randomized control trials meet the "Gold Standard"? A study of the usefulness of RCTs in the What Works Clearinghouse. *American Enterprise Institute*.

Glass, G. V. (2016). One hundred years of research: Prudent aspirations. *Educational Researcher*, *45*(2), 69-72. https://doi.org/10.3102/0013189X16639026

Gonzales, L. D., & Núñez, A. M. (2014). The ranking regime and the production of knowledge: Implications for academia. *Education Policy Analysis Archives*, *22*(31). http://dx.doi.org/10.14507/epaa.v22n31.2014

Good, T., Burross, H., & McCaslin, M. (2005). Comprehensive school reform: A longitudinal study of school improvement in one state. *The Teachers College Record*, *107*(10), 2205-2226. https://doi.org/10.1111/j.1467-9620.2005.00589.x

Gopal, S., & Schorr, L.B. (2016, June 2). Getting "Moneyball" right in the social sciences. *Stanford Social Innovation Review*. Stanford, CA.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement. Abingdon: Routledge*.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8). https://doi.org/10.1371/journal.pmed.0020124

Kraemer, H. C. (2016). Messages for clinicians: Moderators and mediators of treatment outcome in randomized clinical trials. *American Journal of Psychiatry*, *173*(7), 672–679. https://doi.org/10.1176/appi.ajp.2016.15101333

Lipsey, M. W., Puzio, K, Yun, C, Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., . . . Busick, M. D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms*. U.S. Department Of Education. Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncser/pubs/20133000/

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, *36*(2), 127-144. https://doi.org/10.3102/0162373713507480

Plsek, P.E. (1999). Quality improvement methods in clinical medicine. *Pediatrics*, *103*(1 Suppl 203 - 214.

Pogrow, S. (1998). What is an exemplary program and why should anyone care? A reaction to Slavin and Klein. *Educational Researcher*, *27*(7), 22-29. https://doi.org/10.2307/1176057

Pogrow, S. (1999). Rejoinder: Consistent large gains and high levels of achievement are the best measures of program quality: Pogrow responds to Slavin. *Educational Researcher*, *28*(8), 24-26, 31. https://doi.org/10.2307/1176313

Pogrow, S. (2000a). The unsubstantiated "Success" of *Success for All*. Implications for policy, practice, and the soul of the profession. *Phi Delta Kappan*, *81*(8), 596-600. https://doi.org/10.1177/003172170008200114

Pogrow, S. (2000b). *Success for All* does not produce success for students. *Phi Delta Kappan*, *82*(1), 67-80.

Pogrow, S. (2002). *Success for All* is a failure. *Phi Delta Kappan*, *83*(6), 463-468. https://doi.org/10.1177/003172170208300612

Pogrow, S. (2015). *Authentic Quantitative Analysis for Education Leadership Decision-Making and EdD Dissertations: A Practical, Intuitive, and Intelligible Approach*. National Council of Professors of Educational Administration.

Scammacca, N., Vaughn, S., Roberts, G., Wanzek, J., & Torgesen, J. K. (2007). Extensive reading interventions in grades k–3: From research to practice. Portsmouth, NH: *RMC Research Corporation*, Center on Instruction.

Sparks, S. D. (2013, October 30). School improvement model shows promise in first i3 evaluation. *Education Week, 33(11), 8*.

Wanzek, J., Vaughn, S., Scammacca, N. K., Metz, K., Murray, C. S., Roberts, G., & Danielson, L. (2013). Extensive reading interventions for students with reading difficulties after grade 3. *Review of Educational Research*, 163-195. https://doi.org/10.3102/0034654313477212

About the Author

Stanley Pogrow
San Francisco State University; University of Arizona
spogrow@sfsu.edu
http://leadership-quantmethods.blogspot.com/
Stanley Pogrow (PhD, Stanford University) is currently Professor of Educational Leadership and Equity at San Francisco State University, and Professor Emeritus at the University of Arizona. His research and teaching interests are policies and practices for reducing the achievement gap and accelerating the learning of students born into poverty and the use of research to improve leadership decision-making. His most recent book is *Authentic Quantitative Analysis for Education Leadership Decision-Making and EdD Dissertations.*

education policy analysis archives

Volume 25 Number 5 January 23, 2017 ISSN 1068-2341