Moving Behavioral Science From Efficacy to Effectiveness

Denise Hallfors and Hyunsan Cho

Abstract

The gap between scientific knowledge and real world practice continues to be a major conundrum for the behavioral sciences. This paper briefly reviews the development of behavioral research and describes problems that have arisen in meeting the goal of improving behavioral interventions through science. Based on published literature and personal observations, the authors argue that behavioral research has followed too closely after the pharmaceutical research model, with reliance on small efficacy trials under optimal conditions. Specific problems are outlined along with three recommended solutions. In brief, real world feasibility testing is essential, and external validity must become as important as internal validity for evidence of effectiveness.
Keywords: Effectiveness, psychotherapy research, scientist practioner gap, alternative research paradigm.

Introduction

The gap between scientific knowledge and real world practice continues to be a major conundrum for the behavioral sciences. The promise that science can be brought to bear on vexing problems in American society, such as substance abuse and addiction, mental illness, crime and delinquency, obesity, and disease, has prompted an abundance of government initiatives to advance the development and effective use of "evidence-based" interventions (see, e.g., Coalition for Evidenced Based Policy (CEBP), 2002; 2003; Institute of Medicine, 1998; National Institute of Mental Health, 1998; Office of National Drug Control Policy, 1999). The consensus public policy goal is to improve prevention and treatment practice through implementation of interventions found to be effective through rigorous scientific research.

This paper briefly reviews the development of behavioral research and describes problems that have arisen in meeting the consensus goal. We then recommend several remedies for addressing these problems. Although behavioral science has used pharmaceutical research as a prototype, we contend that behavioral research needs to make its own way, a new way grounded more completely in the real world.

Randomized Controlled Trials (RCTs) are considered the "gold standard" for establishing causality and determining the scientific evidence for an intervention's effects. Although the RCT is widely accepted throughout behavioral science today, it is a relatively recent methodological innovation. In the early 1960s, the Food and Drug Administration (FDA) began requiring peer reviewed randomized trials demonstrating a pharmaceutical drug effectiveness before the FDA would allow the drug to be marketed (http://www.fda.gov/cder/about/history/). That policy change, along with parallel support by the National Institutes of Health (NIH) transformed the RCT in medicine from a rare and controversial phenomenon into the final standard for assessing the effectiveness of all new drugs and medical devices (CEBP, 2003). As evidence of impact, the number of clinical research articles based on RCTs surged from about 100 in 1966 to 10,000 in 1995 (Chassin, 1998).

In contrast to medicine, the RCT has been slower to take hold in education, crime, substance-abuse policy and in most areas of social policy, largely because it is more difficult to define and standardize protocols and more complicated to control environmental influences on behavior (CEBP, 2003). Likewise, pharmaceutical and medical research become more complex when applied to patients taking medication

on a regular schedule in their homes, or to changing the way physicians practice medicine. Thus, even medical innovation has suffered from the chasm between what is known through research and what is practiced by physicians and patients (Wells, 1999; Braslow et al., 2005; Tunis et al., 2003).

Nevertheless, behavioral research followed pharmaceutical studies in making the RCT the cornerstone of evidence testing, and in particular, it became the sine qua non for testing efficacy. A seminal paper by Flay (1986) was highly influential in defining necessary stages in the development of public health programs. His linear "phases of research" model followed a logical progression for program testing, including hypothesis development, pilot studies, efficacy trials, effectiveness trials, and dissemination studies.

According to Flay's model, the experimental efficacy trial provides the test of whether a technology, treatment, procedure, or program does more good than harm when delivered under *optimum conditions*. If efficacy could be established, effectiveness trials were to be done next to determine whether a technology, treatment, procedure, intervention, or program does more good than harm when delivered *under real-world conditions*. Flay further differentiated between treatment effectiveness trials, in which implementation fidelity is maintained as much as possible, and implementation effectiveness trials, in which implementation is allowed to vary naturally or by planned comparison. The design for both is large scale experimental or quasi-experimental trials.

More recently, others have defined similar unique characteristics of efficacy versus effectiveness trials. For example, Wells (1999) noted that the efficacy trial optimizes isolation of the treatment effect through design features, such as a control or placebo condition, randomization, standardized treatment protocols, homogeneous samples, and blinding. Wells also noted that the efficacy trial often entails substantial deviations from usual practice conditions by eliminating treatment preferences, providing free care, using specialized providers and settings, maintaining high treatment compliance, and excluding patients with major comorbid conditions. On the other hand, effectiveness studies evaluate effects of interventions under conditions approximating usual care, relying on representative patients and providers in community settings. Cost-effectiveness studies are an essential component and evaluate the marginal difference in outcome for one treatment relative to an alternative. The design is more heterogeneous and quasi-experimental designs are commonly used. Wells and others have thus equated the efficacy trial with greater internal validity, and effectiveness trials with greater external validity (Wells, 1999; Donenberg et al., 1999; Glasgow et al., 2003).

The NIH has followed the general approach of Wells (1999), Flay (1986), and Greenwald and Cullen (1985; authors of a similar paper on cancer research) in program announcements soliciting behavioral studies. For example, the National Institute on Drug Abuse (NIDA) Treatment Branch developed the stage model of behavioral therapies research to conceptualize the transition from initial development of a new treatment intervention to ultimate community utilization (Rounsaville, Carroll, Onken 2001). Stage I consists of initial development and pilot or feasibility testing of new and untested treatments. Stage II consists principally of randomized controlled clinical trials to evaluate efficacy of treatments that have shown promise or efficacy in initial studies. Stage III is intended to address issues of transportability of treatments whose efficacy has been demonstrated in at least two stage II trials (see, e.g., NIH PA-03-06). Akin to specifying the formulation and dosage of medications in FDA standard pharmacotherapy trials, efficacy trials were considered the centerpiece of behavioral therapy development (Rounsaville, Carroll, & Onkin, 2001).

We argue that this emphasis on efficacy trials has increased the gap between research and practice in the behavioral sciences, because few interventions are ever tested beyond this stage. We join other researchers (e.g., Green & Glasgow, 2006; Tunis et al., 2003; Jensen, 2003; Braslow et al., 2005) in attributing much of the gap to an exclusive emphasis on internal validity and a neglect of external validity

in testing interventions. New interventions are produced and tested but they have little salience and poor fit with the world of decision makers. Moreover, government interventions to help close the gap have had very mixed results. In this next section, we describe the major problems.

Problem 1. Too many efficacy trials, too few effectiveness studies

Glasgow and colleagues (2003) argue that although Flay assumed that successful efficacy trials would lead naturally to effectiveness trials, this has not occurred. Instead, scientists developed many small-scale efficacy studies of unknown generalizeability and very few successful effectiveness trials (Glasgow et al., 2002; Oldenburg et al., 2000). For example, Glasgow and colleagues (2002) conducted a comprehensive review of controlled interventions for dietary change, physical activity, and smoking cessation in various healthcare settings (work site, health care, schools, community) that were published in 12 leading health behavior journals between 1996 and 2000. They found an enormous cumulative imbalance in the attention to internal versus external validity.

Other authors have also noted that effectiveness trials (with a focus on external validity) are rare among behavioral interventions, and that few prevention or treatment programs have been tested beyond the small efficacy trials conducted by program developers and implemented under ideal conditions (Greenberg, 2004; Jensen, 2003). In the field of prevention research, programs reporting positive effects in efficacy trials are deemed "model" and marketed as such, usually without independent replication and further testing and development (Hallfors et al., 2006a). Yet evidence from meta-analyses of extant published trials indicates that effects are not robust. Tobler and colleagues found a large drop (from 0.35 to 0.08) in the effect size of school-based interactive programs as the number of students in the study increased (Tobler et al., 2000). Their explanation was that fidelity of implementation decreased when programs went to scale. Lipsey (1992), in a meta-analysis of juvenile delinquency interventions, found that researcher implementation under optimal conditions made the largest single contribution to the R-square change in effect size, adding 0.11.

Problem 2. Efficacy findings may not hold up in effectiveness trials

Independent researchers are often unable to replicate positive efficacy trial findings when testing behavioral interventions in effectiveness trials. Recent examples from prevention science include evaluations of several "Model" programs: *Project Alert* (Ellickson & Bell, 1990; St. Pierre et al., 2006), *Strengthening Families* (Kumpfer et al., 1989; Gottfredson et al., 2006), *Life Skills* (Botvin et al., 1995; Smith et al., 2004), and the *Nurse Home Visitation Program* (Olds et al., 1998; Alper, 2002).

Some effectiveness trials have even found negative effects, that is, the intervention group had worse outcomes than the control group. For example, in our evaluation of *Reconnecting Youth* (Eggert et al., 1994), we found no effects immediately after the program and three negative effects at six months (Hallfors et al., 2006b). These negative effects included greater association with high risk peers, lower association with conventional peers, and fewer prosocial weekend activities. Because the evaluation was a treatment effectiveness trial, implementation fidelity was high (Sanchez et al., 2007) and we concluded that the negative effects were largely due to theory failure. Although the intervention was highly interactive and included cognitive behavioral training to reduce drug use and improve school work, clustering high-risk students in a semester long course led to bonding with deviant peers, and disconnection from conventional peers and activities. This finding is supported by other studies that have found placing high-risk youth in peer group interventions can produce negative effects (Palinkas et al., 1996; Dishion et al., 1999).

The most common explanation of why successful efficacy trial findings fail to be replicated in effectiveness trials is that the quality and fidelity of program implementation are reduced (Elliott & Mihalic 2004; CSAP 2001; Tobler et al., 2000). Program implementation failure (when programs have

not been implemented as designed by the program developer) is referred to as a type III error, and is central to internal validity (Dumas et al., 2001). Efficacy is established based on careful implementation of intervention protocols (fidelity), but almost all behavioral interventions undergo some local adaptation in the real world (Greenberg, 2004; Rogers, 1995). Indeed, adaptations may be so extensive that the program no longer clearly resembles the original evidence-based protocol (Hallfors & Godette, 2002) and may not be effective. From the perspective of community-based participatory approaches (Israel, Eng, Schulz, & Parker, 2005), adaptations are not only pervasive but also generally desirable. Given the likelihood of adaptations in the real world, some reasonable accommodations are needed to render the intervention feasible and effective.

Another problem is the quality of reporting and analysis of RCT data. The CONSORT group (Altman et al., 2001) determined that there was overwhelming evidence of inadequate reporting and design associated with biased estimates of medical treatment effects. A related problem is the tendency for researchers to submit and editors to publish articles with positive findings but not those with null or negative findings. This publication bias or "file-drawer effect" (Rosenthal, 1979) has been confirmed in most areas, such as healthcare, psychology, education, and behavioral research (Dickersin 2002; Lipsey and Wilson 1993; Torgerson 2006). Publication bias is a major problem in assessing the evidence for intervention effectiveness.

Problem 3. Behavioral interventions are often not feasible.
Behavioral interventions tested in efficacy trials are often not feasible in real world settings under usual conditions. Such interventions are tested and marketed by developers who may not see the "rubbing points" that make them unacceptable in the real world. As an example, we tested the *Reconnecting Youth* program (Eggert et al., 1994) for high risk high school students (Hallfors et al., 2006b;2006c). Imbedded in the evaluation survey was a screen for suicide risk which was considered a necessary complement to the program (personal communication, Leona Eggert, 2001). Using the screen, we flagged almost 30% of students as at risk for suicide, necessitating a follow up assessment. Follow-up interviews found that approximately 80% were false positives, which resulted in considerable resistance from school counselors to conduct the screen and the follow-up, and the screen was eventually dropped from the survey (see Hallfors et al., 2006c).

Similarly, Gottfredson and colleagues (2006) found that the Strengthening Families Program (SFP) showed low feasibility for families and agencies that would be expected to adopt the program. SFP was originally tested and found effective in reducing parent, family and youth risk factors for substance use and later youth substance use in children of drug abusers in treatment (Kumpfer & DeMarsh, 1985). SFP was subsequently identified as an effective prevention program by several federal agencies interested in reducing substance use and delinquency. However, an effectiveness trial targeting predominantly African American, urban populations found minimal effects on child outcomes (Gottfredson et al., 2006). The enormous difficulties related to recruitment and retention of families for the study were evidence that the program was either not acceptable or not a high priority for many clients. In addition, the extremely high turnover rate among trainers and site coordinators indicated that much greater resources than expected were needed to administer the program with fidelity.

Problem 4. Government involvement: Blessing or bane?
Government can play a critical role in the transfer of research into practice by reviewing evidence and designating rigorously tested interventions as research-based, and by requiring organizations funded with government dollars to select research-based programs. However, government involvement can be a two-edge sword, creating false incentives that result in unintended consequences. For example, diffusion theory indicates that incentives can increase the rate of adoption of new practices and motivate individuals and organizations that would otherwise not adopt, but commitment to the adoption decision may be low, limiting the intended consequences of adoption (Rogers, 1995). As an example, we

evaluated the impact of a federal education policy requiring school districts to adopt research-based drug and violence prevention programs in order to maintain school funding (Hallfors & Godette, 2002). Although 59% of school districts reported using one or more research-based programs, only 19% were implementing the programs as designed.

Some government agencies have either commissioned groups to review evidence for effective interventions, or have taken on the task themselves. For example, in order to determine what qualified as "research-based" or "evidence-based," the Substance Abuse and Mental Health Services Administration (SAMHSA) established the National Registry of Evidence-based Programs and Practices (NREPP). Between 1997 and 2004, SAMHSA reviewed and rated more than 1,100 programs, designating more than 150 as Promising, Effective, or Model programs (US DHHS, 2005). In 1998, the Department of Education commissioned an expert panel to establish criteria for Promising and Exemplary programs, review applications, and make recommendations to the Secretary; their report was released in 2001 (Petrosino, 2003; Safe, Disciplined, and Drug-Free Schools Expert Panel, 2001). In general, both agencies deemed programs as "Promising" if they met some lesser threshold of evidence than Effective or Exemplary programs. NREPP further distinguished between Effective and "Model" programs by the ability of developers to disseminate and support program implementation by end-users. Other federal agencies and private organizations have likewise developed "lists" to guide schools in choosing "evidence-based" programs.

The NREPP list is currently the most influential of all of the school-based prevention program lists (Hallfors, et al. 2006a). Because of its influence in determining which prevention programs would be selected by schools, it was subject to intense lobbying by vendors of such programs. NREPP collects its data by soliciting program developers, asking them to provide study findings as evidence for their interventions. This data collection method results in a systematic bias in favor of proprietary programs (rather than policy or other structural interventions), raises small efficacy trials to the level of "evidence," and omits effectiveness trials led by independent evaluators.

NREPP faced many criticisms about its review process and determination of whether a program was "Promising," "Effective," or "Model" (US DHHS, 2005; Hallfors et al., 2006a). However, their solution was simply to provide scoring information from multiple dimensions for every program reviewed, leaving decisions about evidence to the NREPP user (US DHHS, 2006). Based on our survey of state Safe and Drug Free Schools Directors, we concluded that this will lead to the widespread assumption that all programs on the NREPP list are sufficiently "evidence-based" even when programs with very weak evidence are included on the list (Hallfors et al., 2006a). This is troubling since, under the new NREPP procedures, a program would qualify for inclusion if the developer provided a single study using a single group pre- to post-test design that showed just one positive behavioral outcome with significant change (US DHHS, 2006).

Solution 1. Improve the development of behavioral interventions

Evidence across many disciplines points to the snail slow transfer of research-based interventions to practice. One implication is that novel methods for intervention development and testing approaches are needed that maximize both internal and external validity. The NIH, as the predominant funding source of intervention development, obviously plays a pivotal role. However, these issues are conceptualized and addressed in somewhat different ways across and within the Institutes. For example, the Prevention Branch at NIDA conceptualizes the essential development steps as basic research (theory and hypothesis development) leading to efficacy research, leading to small-scale effectiveness research, leading to large-scale effectiveness research (personal communication, Dr. Elizabeth Robertson, February 2007; see Figure 1). Systems research is a subsequent step to address the transfer of effective interventions into

widespread practice. Services research questions (what works? under what circumstances? for which conditions? at what cost?) are central to this model, ideally even at the basic research and efficacy research stage. This type of approach implies that behavioral science needs to shed light not only on an intervention's efficacy but also on how well efficacious interventions can actually work in diverse settings, provider and patient populations, and practice circumstances (Braslow et al., 2005). It implies the importance of ascertaining feasibility in transferring interventions to real world settings.

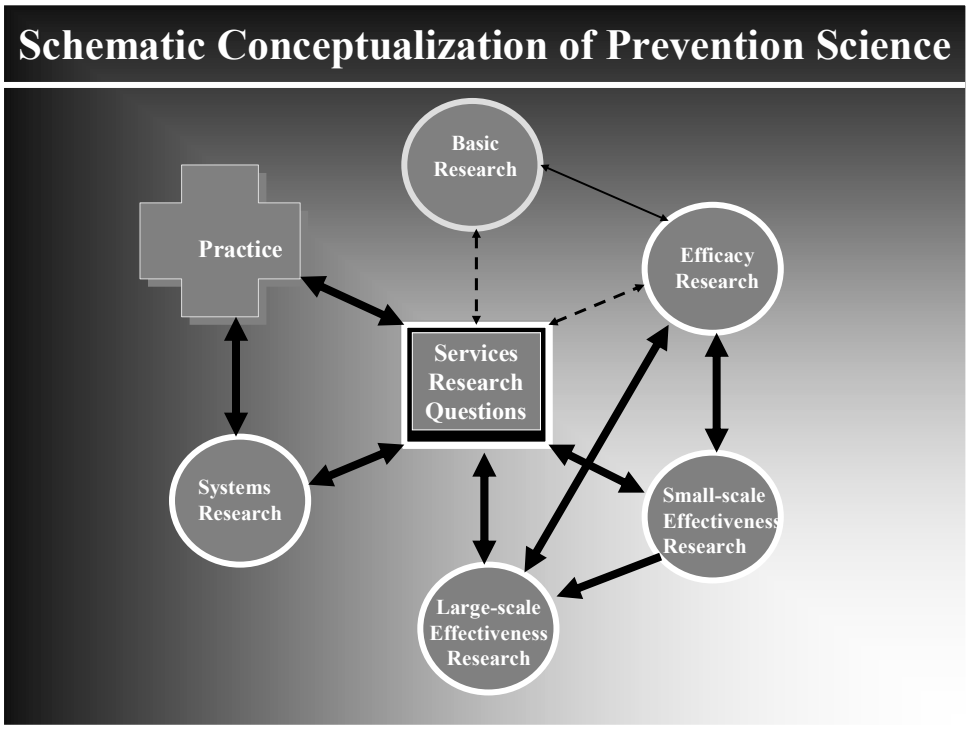

Figure 1: Schematic Conceptualization of Prevention Science

Since efficacy research is, by definition, conducted under optimal conditions, a preponderance of evidence gained solely from efficacy trials may be one reason for the lack of transfer to real world settings. Interventions that are efficacious under optimal conditions may not be acceptable to real world decision makers, and they may not be feasible or effective. An alternative to efficacy trials would be to conduct small controlled pilot trials to test feasibility and examine effects. If effects are promising, and if the intervention is acceptable and feasible (as judged by real world implementers and an advisory panel of stake holders), then the intervention should be documented in a manual for further testing. The National Institute of Mental Health (NIMH) encourages researchers to use the R34 mechanism for this purpose (see PAR-06-248).

The pilot stage could take the place of the efficacy trial for behavioral interventions. This would be an improvement because the pilot would require the developer to document feasibility issues, and data would not be allowed to rise to the level of "evidence" as commonly occurs with efficacy trials (Hallfors et al., 2006a). Instead, the first stage of "evidence" would be a larger hybrid trial that maximizes both internal and external validity features. This type of trial would include a control condition, randomization (preferably at the aggregate level), monitoring of treatment delivery, and training real world providers, practicing in typical community treatment settings, with typical participants, testing outcomes over a longer period of time (e.g., 1-2 years), and assessing feasibility and cost.

Our one nod to testing under somewhat optimal conditions is to select settings that are relatively "innovative" as defined by diffusion of innovation theory (Rogers, 1995). Rogers (1995) describes five different adopter categories: innovators, early adopters, early majority, late majority, and laggards. Efficacy trials are often conducted in "innovator" settings. These organizations are intensely interested in new ideas and willing to take a risk before any of their peers; however, they represent a very small minority (according to Rogers, about 2.5%). Our favorite participants fall within the next two categories. Early adopters are well respected by peers and are watched closely when they adopt new practices. They make up about 13.5% of adopters. Early majority adopters comprise about 34%, and they are noted for being deliberate and careful in accepting new ideas. Our experience in school settings indicates that organizations in these two categories are well run, more likely to follow the research protocols, and less likely to drop out of the study than the last two groups, and they are also more influential than innovators (Hallfors et al., 2006b). Late majority and laggard organizations will join under pressure, but they are skeptical of change, prefer their usual way of doing things, and do not make good research participants.

The hybrid model (Wells, 1999; Roy-Byrne et al., 2003; Carroll & Rounsaville., 2003) is analogous to Flay (1986)'s Treatment Effectiveness Trial. Our evaluation of RY followed the treatment effectiveness model, and was a randomized controlled trial that closely monitored program implementation according to the developer's manual, using typical staff and participants, in typical urban settings (Cho et al., 2005; Hallfors et al., 2006b; Sanchez et al., 2007). To emphasize the need for standardized reporting protocols regarding external validity on these types of trials, we believe that the RE-AIM (reach, effectiveness, adoption, implementation, maintenance and cost) assessments and reporting (see Glasgow et al., 2003) would help make this information more useful to decision makers about whether an intervention would be appropriate and cost effective for participants in their setting.

In terms of how to move from efficacy to effectiveness, we contend that the hybrid model could replace at least some behavioral efficacy trials. Given the high cost of clinical trials, we believe that smaller pilot studies, (e.g., using the R01, R21 or R34 NIH mechanisms), could assess feasibility and pilot efficacy using small randomized or well-matched controlled trials. If pilot findings are promising, the researchers could manualize the intervention and conduct a hybrid trial that maximizes both internal and external validity (perhaps analogous to small-scale effectiveness research in NIDA's model). If findings show that the intervention is effective under real world conditions, the next step should be a replication of the hybrid trial with an independent research team.

Replication of positive effects by an independent team of scientists should be the gold standard of intervention effectiveness. Independent scientists are much more likely than developers to approach an intervention like the typical end-user, since developers are extremely familiar with the intervention and deeply vested in outcomes. At the same time, independent scientists have greater resources to replicate and evaluate the intervention with fidelity in the field than typical end-users. If scientists are not able to replicate positive results in an RCT, then it is unlikely that widespread adoption will result in improved outcomes. This method of intervention development would greatly strengthen the evidence from research. It would then make sense for government to provide incentives for organizations to change their policies and practice and adopt interventions that hold up to this level of testing.

Solution 2. Improve the relevance of research to practice

A second solution is to collaborate with clinicians and practitioners and evaluate the interventions that they see as most important and acceptable to their practice. Evidence-based interventions are often seen as top-down impositions by the practitioners who are expected to adopt and sustain them. This second solution suggests a "bottom-up" approach in which practitioners are considered the experts. The NIMH document "Bridging Science to Service" (1998) included the recommendation that "NIMH should support research to identify common practices believed to be helpful and bring them under research

scrutiny, that is, ascertain what is going on in the practice community and determine how much of that is beneficial." (page 5). The NIDA Prevention Branch is also interested in evaluating prevention interventions that have been widely implemented but never evaluated (personal communication, Elizabeth Robertson, February 2007).

The importance of evaluating widely used programs relates to the reach and relevance of such findings. Three examples of widespread school-based prevention interventions that have not been evaluated include school resource officers, student assistance programs, and alternative schools. A survey of over 100 school districts in 10 states (Hallfors et al., 2000; 2001) found that 72% of schools had, or were planning to have, Student Assistance Programs at the high school level; 69% at the middle school level. Almost 50% of school districts reported that school resource officers were used to a "great extent" in the district's substance abuse prevention program. Alternative schools are also widespread, but under-studied (Gottfredson, 2001). Systematic studies are needed to examine how these programs are being implemented, what impact they are having on preventing drug use and violence, and at what cost. This type of research would be of immediate use to principals and other decision makers within schools.

Because popular programs are already in widespread use, testing new interventions compared to those already in place represents a better way to help decision makers in their real world choices. Tunis, Stryer, and Clancy (2003) similarly argued that many medical decision makers, policy makers, clinicians, and consumers do not find much of the evidence base from highly controlled randomized efficacy trials to be very relevant to their situation or the concerns that they have. These scientists recommend conducting "practical trials" that assess outcomes important to decision makers, such as cost-effectiveness and quality of life, and using representative (or at least heterogeneous) samples of patients and settings. They also recommend evaluating new pharmaceuticals against popular treatments rather than no treatments or placebo controls. These practical trials, designed to make a decision about which treatment to use, are contrasted with "explanatory" trials, designed to detect an effect in a new treatment. Practical trials elevate the practice perspective, thus making the research contribution more relevant.

Another way to bridge the gap is to bring researchers into practice settings as consultants to clinicians. New and innovative ideas from clinicians can be solicited and these "front-line" practitioners can also engage in generating clinical evidence. Sullivan and colleagues (2005) describes a novel method for involving clinicians in the development and testing of interventions within the Veterans Administration System. Clinicians applied for funding to test their intervention ideas and collaborated with research scientists in evaluating these ideas. This approach can help clinicians become intimately acquainted with the generation of evidence and the testing of ideas about "what works."

Solution 3. Registering findings for systematic meta-analyses

The third solution entails the gathering of extant evidence on behavioral interventions to better use the knowledge base. In a recent program announcement (PAR-06-039), NIH recognized that each year, billions of U.S. tax dollars are spent on research and hundreds of billions are spent on service delivery programs. It goes on to say that, despite this investment, little is known about how to systematically disseminate lessons learned from research and practice to improve the care provided to people in this country.

Surprisingly, there is no current requirement to report findings from all NIH funded behavioral trials in a standardized way. Yet this seems a fundamental first step for assessing and using the knowledge base. We recommend that principal investigators from all behavioral trials be required to report findings to a central data base or registry, using a standardized reporting form that will support meta-analyses of findings.

One way to standardize reporting would be to use the Consolidated Standards of Reporting Trials (CONSORT) (http://www.consort-statement.org). The objective of the CONSORT is to facilitate critical appraisal and interpretation of RCTs by providing guidance to authors about how to improve the reporting of their trials (Altman et al., 2001). CONSORT reporting has been supported by many journals including the Lancet, British Medical Journal, Journal of the American Medical Association, and Annals of Internal Medicine, and a growing number of biomedical editorial groups. The Evidence-Based Behavioral Medicine Committee of the Society of Behavioral Medicine researchers has recommended adoption of a modified version of the CONSORT criteria for reporting randomized controlled trials (Glasgow et al., 2004). In order to improve reporting of external validity, their adapted version includes seven elements from the RE-AIM framework (Glasgow et al., 2003).

This type of registry is complex but we now have considerable experience in developing and using clinical trial data from the ongoing work of the Cochrane Collaboration for medical research (http://www.cochrane.org/) and the newer Campbell Collaboration for social, educational, and behavioral interventions (http://www.campbellcollaboration.org/). Systematic reviews of RCTs and other methodologically rigorous trials are essential if we are to reap the benefits of the massive behavioral science investments. Registering the methods and findings of all trials is an essential step towards solving the publication bias that has made some ineffective interventions appear to be "evidence-based." Registries can also provide an inventory of studies to help funding agencies generate new research to fill in gaps. Green & Glasgow (2006) go further and recommend that registries or repositories of evaluations conducted more routinely in more representative settings and populations could further strengthen the external validity of evaluation literature on interventions.

The NIDA Prevention Branch is currently gathering evidence about how state and local systems select, plan, and implement evidence-based substance abuse prevention interventions and whether these efforts show effects at the community or state level. NIDA currently provides approximately one million dollars annually to the Substance Abuse and Mental Health Services Agency (SAMHSA) to evaluate the massive Strategic Prevention Framework State Incentive Grant (SPF-SIG; http://prevention.samhsa.gov/grants/sig.aspx). SAMHSA has awarded funding to some 37 states, territories, and tribal governments to determine whether requiring states and communities to conduct needs assessments, adopt goals, select evidence-based interventions, and evaluate progress towards goals using epidemiological data can reduce substance abuse and negative consequences stemming from substance abuse. Although implementation of the SPF-SIG at the community level will be evaluated as a "package" consisting of multiple activities, additional analyses are expected to help identify specific features of SPF implementation, including the specific interventions used, that were particularly effective.

This also holds the promise of providing more comprehensive information about how and whether adaptations to specific interventions are effective when used locally. Adaptation data can help illuminate issues related to the balance between fidelity of implementation and local customization. Green and Glasgow (2006) have suggested that the solution may lie in the specification and documentation of: 1) a limited set of key components or principles underlying an efficacious intervention; 2) the range of permissible adaptations that still retains the essential elements; and 3) justification of theory-driven and experience-driven deviations from the tested intervention. Empirical information from the RCT registry and SPF-SIG data base may help determine key components and the range of permissible adaptations that preserve positive outcomes.

Summary

There is a consensus goal among scientists, practitioners, consumers, and policy makers to improve behavioral interventions through scientific knowledge. But despite growing consensus, heavy investment, and strong advancements in rigorous methodology, this goal remains elusive. We argue that

behavioral research has followed too closely after the pharmaceutical and medical product research model, with reliance on small efficacy trials under optimal conditions. While efficacy trials may be appropriate for medical product testing, they are not the best method for behavioral intervention research. Real world feasibility testing is essential, and external validity must become as important as internal validity for evidence of effectiveness.

We outlined four main problems related to reaching the goal of improved practice through scientific research: 1) efficacy trials are rarely followed by effectiveness trials; 2) when they are, they often show null outcomes; 3) as well as feasibility problems; and 4) government policies trying to speed the transfer of interventions prior to adequate effectiveness testing, have resulted in low quality adoption. We then outlined three solutions to address these problems: 1) increasing external validity while maintaining high internal validity through pilot studies followed by "hybrid" effectiveness trials and independent researcher replications; 2) increasing the weight of collaborations with practitioners, encouraging "bottom up" evaluations; and 3) registering findings from all randomized or well-controlled intervention trials and expanding the research base with other evidence from field testing.

In this paper, we add our voice to those of Glasgow, Tunis, and many others who have suggested new strategies to meet the consensus goal. We argue that behavioral research must separate from the pharmaceutical prototype and chart a new course that will require feasibility testing and external, as well as internal, validity. In this way, behavioral science can finally make significant progress towards closing the gap and improving human health and well being.

References

Alper, J. (2002). The nurse home visitation program. In S.L. Issacs & J.R. Knickman (Eds.), *To Improve Health and Health Care. Volume V. The Robert Wood Johnson Anthology* (pp. 3-22). an Francisco: Jossey Bass.

Altman, D.G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med, 134*(8), 663-694.

Botvin, G.J., Baker, E., Dusenbury, L., Botvin, E.M., & Diaz, T. (1995). Long-Term Followup Results of a Randomized Drug Abuse Prevention Trial in a White, Middle-Class Population. *Journal of the American Medical Association, 273*(14), 1106-1112.

Braslow, J.T., Duan, N., Starks, S.L., Polo, A., Bromley, E., & Wells, K B. (2005). Generalizability of studies on mental health treatment and outcomes, 1981 to 1996. *Psychiatr Serv, 56*(10), 1261-1268.

Campbell Collaboration. (n.d.). C2 Home: What helps? What harms? Based on what evidence? Retrieved from http://www.campbellcollaboration.org/.

Carroll, K.M. & Rounsaville, B.J. (2003) Bridging the gap: a hybrid model to link efficacy and effectiveness research in substance abuse treatment. *Psychiatr Serv*, 54, 333-9.

Center for Substance Abuse Prevention (CSAP). (2001). *Finding the Balance: Program Fidelity and Adapation in Substance Abuse Prevention. A State-of-the-Art Review.* Washington, DC: US Department of Health and Human Services.

Center for Substance Abuse Prevention (CSAP). (n.d.). State Incentive Grants (SIG). Retrieved from http://prevention.samhsa.gov/grants/sig.aspx.

Chassin, M.R. (1998). Is Health Care Ready for Six Sigma Quality? *Millbank Quarterly*, 76(4), 574.

Cho, H., Hallfors, D., Sanchez, V. (2005) Evaluation of a High School Peer Group Intervention for At-Risk Youth. *Journal of Abnormal Child Psychology,* 33(3), 363-374.

Coalition for Evidence-Based Policy (CEBP). (2002). *Bringing evidence-driven progress to education.* Washington, DC.

Coalition for Evidence-Based Policy (CEBP). (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide.* Washington, DC.

Cochrane Collaboration. (n.d.). The reliable source of evidence in health care. Retrieved from http://www.cochrane.org/.

CONSORT. (n.d.). CONSORT, Strength in Science, Sound Ethic. Retrieved from http://www.consort-statement.org/.

Dickersin, K. (2002). Reducing reporting biases. In I. Chalmers, I. Miline and U. Trohler (Eds) *The James Lind Library* (www.jameslindlibrary.org)

Dishion, T.J., McCord, J., & Poulin, F. (1999). When Interventions Harm: Peer Groups and Problem Behavior. *American Psychologist, 54*(9), 755-764.

Donenberg, G.R., Lyons, J.S. & Howard, K.I. (1999). Clinical trials versus mental health services research: Contributions and connections. *Journal of Clinical Psychology*, 55, 1135-1146.

Dumas, J.E., Lynch, A.M., Laughlin, J.E., Phillips Smith, E., Prinz, R.J. (2001). Promoting Intervention Fidelity. Conceptual Issues, Methods, and Preliminary Results from the Early Alliance Prevention Trial. *American Journal of Preventive Medicine*, 20, 38-47.

Eggert, L.L., Thompson, E.A., Herting, J.R., Nicholas, L.J., Dicker, B.G. (1994). Preventing adolescent drug abuse andhigh-school dropout through an intensive school-based social network development program. *American Journal of Health Promotion*, 4(8), 202–215.

Ellickson, P.L., & Bell, R.M. (1990). Drug Prevention in Junior High: A Multi-Site Longitudinal Test. *Science, 247*, 1299-1305.

Elliot, D.S. and Mihalic, S. (2004). Issues in Disseminating and Replicating Effective Prevention Programs. *Prevention Science*, 5(1), 47-53.

FDA History Office. (n.d.). A Brief History of the Center for Drug Evaluation and Research. Retrieved January, 2007 from http://www.fda.gov/cder/about/history/.

Flay, B. R. (1986). Efficacy and Effectiveness Trials (and other phases of research) in the Development of Health Promotion Programs. *Preventive Medicine*, 14, 451-474.

Glasgow, R.E., Bull, S.S., Gillette, C., Klesges, L.M. & Dzewaltowski, D.A. (2002). Behavior change intervention research in healthcare settings - A review of recent reports with emphasis on external validity. *American Journal of Preventive Medicine*, 23, 62-69.

Glasgow, R.E., Klesges, L.M., Dzewaltowski, D.A., Bull, S.S., Estabrooks, P. (2004). The future of health behavior change research: What is needed to improve translation of research into health promotion practice? *Annals of Behavioral Medicine, 27*(1), 3-12.

Glasgow, R.E., Lichtenstein, E., Marcus, A.C. (2003). Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. *American Journal of Public Health*, 93(8), 1261-1267.

Gottfredson, D. (2001*). Schools and Delinquency*. Cambridge University Press, Cambridge, U.K.

Gottfredson, D., Kumpfer, K., Polizzi-Fox, D., Wilson, D., Puryear, V., Beatty, P., et al. (2006). The Strengthening Washington DC Families project: A randomized effectiveness trial of family-based prevention. *Prevention Science, 7*(1), 57-74.

Green, L.W. & Glasgow, R.E. (2006) Evaluating the relevance, generalization, and applicability of research - Issues in external validation and translation methodology. *Evaluation & the Health Professions*, 29, 126-153.

Greenberg, M.T. (2004) Current and future challenges in school-based prevention: The researcher perspective. *Prevention Science*, 5, 5-13.

Greenwald, P., & Cullen, J.W. (1985). The new emphasis in cancer control. *J Natl Cancer Inst, 74*(3), 543-551.

Hallfors, D., Brodish, P., Khatapoush, S., Sanchez, V., Cho, H. (2006c) Feasibility of Screening Adolescents for Suicide Risk in 'Real World' High School Settings. *American Journal of Public Health*, 96(2), 282-287.

Hallfors, D., Cho, H., Sanchez, V., Khatapoush, S., Kim, H. M., & Bauer, D. (2006b). Efficacy vs effectiveness trial results of an indicated "model" substance abuse program: Implications for public health. *American Journal of Public Health, 96*(12), 2254-2259.

Hallfors, D., Godette, D. (2002). Will the "Principles of Effectiveness" Improve Prevention Practice? Early Findings from a Diffusion Study. *Health Education Research,* 17(4), 461-470.

Hallfors, D., Pankratz, M., Hartman, S. (2006a). Does Federal Policy Support the Use of Scientific Evidence in School-Based Prevention Programs? *Prevention Science*, published online, December 13, 2006, doi: 10.1007/s11121-006-0058-x.

Hallfors, D., Pankratz, M., Sporer, A. (2001). Drug Free Schools Survey II: Report of Results. A report to the Robert Wood Johnson Foundation.

Hallfors, D., Sporer, A., Pankratz, M., Godette, D. (2000). Drug Free Schools Survey: Report of Results. A report to the Robert Wood Johnson Foundation.

Institute of Medicine. (1998). *Bridging the Gap Between Practice and Research: Forging Partnerships with Community-Based Drug and Alcohol Treatment*. Washington, DC: National Academy Press.

Israel, B.A., Eng, E., Schulz, A.J., Parker, E.A. (2005). *Methods in community-based participatory research for health.* San Francisco: Jossey-Bass.

Jensen, P.S. (2003). Commentary: the next generation is overdue. *J Am Acad Child Adolesc Psychiatry, 42*(5), 527-530.

Kumpfer, K.L., DeMarsh, J.P. (1985). Prevention of chemical dependency in children of alcohol and drug abusers. *NIDA Notes*, 5, 2-3.

Kumpfer, K.L., DeMarsh, J.P., Child, W. (1989). Strengthening Families Program: Children's skill training curriculum manual (prevention services to children of substance-abusing parents). Salt Lake City, UT: Department of Health, Alcohol and Drug Research Center.

Lipsey, M.W. (1992) Juvenile Delinquency Treatment: A Meta-Analytic Inquiry into the Variability of Effects, in: Cook, T. D., Cooper, H., Cordray, D. S. et al. (Eds.) *Meta-Analysis for Explanation: A Casebook* (pp. 83-127). New York: Russell Sage Foundation.

Lipsey, M.W. and Wilson, D.B. (1993). The Efficacy of Psychological, Educational and Behavioural Treatment: Confirmation from Meta-analysis. *American Psychologist*, 12, 1181-1209.

National Institute of Mental Health. (1998). *Bridging Science and Service: A Report by the National Advisory Mental Health Council's Clinical Treatment and Services Research Workgroup*. Washington, DC. Available at www.nimh.nih.gov/publicat/nimhbridge.pdf

Office of National Drug Control Policy. (1999). *National drug control strategy-Performance measures of effectiveness: Implementation and findings*. Washington, DC: Executive office of the President.

Oldenburg, B.F., French, M.L. & Sallis, J.F. (2000) Health behavior research: The quality of the evidence base. *American Journal of Health Promotion*, 14, 253-257.

Olds, D., Henderson, C.R., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., et al. (1998). Long-term effects of Nurse Home Visitation on Children's Criminal and Antisocial Behavior: 15-year Follow-up of a Randomized Controlled Trial. *Journal of the American Medical Association, 280*(14), 1238-1244.

Palinkas, L.A., Atkins, C.J., Miller, C., & Ferreira, D. (1996). Social skills training for drug prevention in high-risk female adolescents. *Prev Med, 25*(6), 692-701.

Petrosino, A. (2003). Standards for evidence and evidence for standards: the case of school-based drug prevention. *Ann Am Acad Pol Soc Sci*, 587, 180-207.

Rogers, E. M. (1995). *Diffusion of Innovations*. New York, NY: Free Press.

Rosenthal, R. (1979). The ''file drawer problem'' and tolerance for null results. *Psychol Bull,* 86, 638–641.

Rounsaville, B.J., Carroll, K.M. & Onken, L.S. (2001). A Stage Model of Behavioral Therapies research: Getting started and moving on from stage I. *Clinical Psychology-Science and Practice*, 8, 133-142.

Roy-Byrne, P.P., Sherbourne, C.D., Craske, M.G., Stein, M.B., Katon, W., Sullivan, G., et al. (2003). Moving treatment research from clinical trials to the real world. *Psychiatric Services, 54*(3), 327-332.

Safe, Disciplined, and Drug-Free Schools Expert Panel. (2001). Exemplary Programs. Available at: http://www.ed.gov/offices/OERI/ORAD/KAD/expert_panel/2001exemplary_sddfs.html and http://www.ed.gov/admins/lead/safety/exemplary01/panel.html.

Sanchez, V., Steckler, A., Nitirat, P., Hallfors, D., Cho, H., Brodish, P. (2007) Fidelity of Implementation in a Treatment Effectiveness Trial of Reconnecting Youth. *Health Education Research, 22, 95-107.*

Smith, E.A., Swisher, J.D., Vicary, J.R., Bechtel, L.J., Minner, D., Henry, K.L., et al. (2004). Evaluation of Life Skills Training and Infused-Life Skills Training in a rural setting: Outcomes at two years. *Journal of Alcohol and Drug Evaluation*, June, 51-70.

St. Pierre, T.L., Osgood, D.W., Mincemoyer, C.C., Kaltreider, D.L., Kauh, T.J. (2006). Results of an Independent Evaluation of Project ALERT Delivered in Schools by Cooperative Extension. *Prevention Science*, Jan 2006, 1-13.

Sullivan, G., Duan, N., Mukherjee, S., Kirchner, J., Perry, D., & Henderson, K. (2005). The role of services researchers in facilitating intervention research. *Psychiatric Services, 56*(5), 537-542.

Tobler, N.S., Roona, M.R., Ochshorn, P., Marshall1, D.G., Streke1, A.V., Stackpole, K.M. (2000). School-Based Adolescent Drug Prevention Programs: 1998 Meta-Analysis. *Journal of Primary Prevention*, 20(4), 275-336.

Torgerson, C.J. (2006) Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies*, 54, 89-102.

Tunis, S.R., Stryer, D.B. & Clancy, C. M. (2003) Practical clinical trials - Increasing the value of clinical research for decision making in clinical and health policy. *Journal of the American Medical Association*, 290, 1624-1632.

US Department of Health and Human Services (DHHS). (2005). Notice: Request for Comments: National Registry of Evidence-Based Programs and Practices (NREPP). *Federal Register*, 70(165).

US Department of Health and Human Services (DHHS) (2006). Changes to the National Registry of Evidence-Based Programs and Practices (NREPP). *Federal Register*, 71(49), 13132–13155.

Wells, K.B. (1999). Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *American Journal of Psychiatry,* 156, 5–10.

Author Contact Information:

Denise Hallfors
The Pacific Institute for Research & Evaluation
1516 E. Franklin St., Suite 200
Chapel Hill, NC 27514
Phone: 919-265-2612
Fax: 919-265-2659
Email: hallfors@pire.org

Hyunsan Cho
The Pacific Institute for Research & Evaluation
1516 E. Franklin St., Suite 200
Chapel Hill, NC 27514
Phone: 919-265-2620
Fax: 919-265-2659
Email: cho@pire.org
