Check for updates

Evaluation & the Health Professions
1-34
© The Author(s) 2018
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0163278718772886
journals.sagepub.com/home/ehp

Strategies to Avoid Replication Failure With Evidence-Based Prevention Interventions: Case Examples From the Strengthening Families Program

Karol L. Kumpfer[1], Lawrence M. Scheier[2], and Jaynie Brown[3]

Abstract

Research has found disturbing long-term effects of poor parenting on children's behavioral health including addiction, delinquency, depression/anxiety, and poorer health as adults. Poor parenting practices thus contribute substantially to the health crisis in America. However, skilled, nurturing parents, or caretakers can help youth avoid these developmental problems. A number of family and parenting evidence-based interventions

[1]Strengthening Families Program, LLC, Salt Lake City, UT, USA
[2]LARS Research Institute, Inc., Scottsdale, AZ, USA
[3]Strengthening Families Program, Salt Lake City, UT, USA

Corresponding Author:
Karol L. Kumpfer, Strengthening Families Program, LLC, 817 East 17th Avenue, Salt Lake City, UT 84103, USA.
Email: kkumpfer@xmission.com

(EBIs) that teach parenting skills are now available for dissemination. Unfortunately, replications of EBIs do not always produce the original positive results. Organizations that seek to use family EBIs to improve parenting and family skills need to avoid practices that create replication failure. We examine several possible factors that contribute to replication failure using examples from five replications of the EBI "Iowa Strengthening Families Program for ages 10–14." We then share six strategies conducive to avoid replication failures including (1) choosing the right program and implementation strategy for the population, (2) administering the right "dosage," (3) choosing and properly training implementers, (4) maintaining program integrity and adherence, (5) ensuring cultural sensitivity, and (6) ensuring accurate and complete reporting of evaluation results. These guidelines can advance prevention science to meet the demands of a growing public health agenda.

Keywords
family-based programs, evaluation, replication failure, implementation, program fidelity

There is now concrete evidence from long-term randomized control trials (RCTs) that family evidence-based interventions (EBIs) can produce upward of 50% reductions in various behavioral health disorders (e.g., Dishion et al., 2008) and continued evidence that strong, resilient families can avoid adverse developmental outcomes (e.g., Kumpfer, Magalhaes, & Xie, 2017). Notwithstanding, we still lack hard and fast rules about implementing EBIs at the ground level where agencies and organizations face a crisis in fostering the health and well-being of America. Indeed, and despite their very best efforts at taking family-based programs to scale, there has recently been a spate of *replication failures* reported in the literature. Replication failure, one of the several translational science challenges, arises when EBIs disseminated and evaluated in tightly controlled efficacy trials are later delivered by independent agencies and research teams under less controlled effectiveness trial conditions and are unable to replicate the original efficacy trial outcomes.

This compendium of negative or null findings may be partly responsible for the *crisis in replication* that is afflicting psychology (Maxwell, Lau, & Howard, 2015) and that extends to prevention science (Valentine et al., 2011). Maxwell, et al. (2015) cite several major reasons for

replication failure as investigators seeking to control Type I error, including but not limited to using different procedures and following different protocols, unique samples, varying measurements, and low powered studies. To satisfy statistical requirements, replications require a much larger sample than the original study, which rarely occurs. Also sampling variability in effect sizes can lead to underpowered replication samples. Confidence intervals may be superior to point estimates of effect sizes in this instance. Any number of factors can contribute to Type II errors (concluding an effective intervention does not work) including procedural variation in implementation and evaluation errors. Replication failure can also include omissions by the researcher to report on all the facets of family change documented in the intervention—and selectively focus on only one or two outcomes.

In this article, we address several issues of paramount importance that we believe are related to replication failure as it specifically affects family-based parenting skills and youth drug prevention programs. We first introduce readers to a brief history of the Strengthening Families Program (SFP), a family-based parenting skills training program that targets a wide range of mental and behavioral disorders in children and youth including drug use. The program is generally administered through community agencies and can also involve the participation of schools, clinics, and various community partnerships. It is for this reason that it provides an excellent backdrop against which we can dissect the many reasons for replication failure including carefully examining implementation concerns and the myriad of ways problems in program delivery can affect program outcomes.

We then share six program-based strategies that agencies and evaluators can engage to avoid replication failures.[1] The six strategies we feel deserve the most attention include (1) matching program type to target audience (i.e., implementing the universal seven-session SFP with high risk families when the selective 10 or 14 weeks SFP likely would have worked better); (2) administering the right dosage (i.e., avoiding cutting dosage by eliminating boosters); (3) properly training implementers that are invested in the program goals (i.e., ensuring buy-in and high implementation enthusiasm); (4) maintaining program integrity by keeping intact core components (i.e., including all key active ingredients) and adhering to the program delivery format (i.e., keeping whole families intact for trainings); (5) culturally tailoring the program (i.e., considering different cultural mores that influence family dynamics); and (6) using the right evaluation methodology (i.e., design, assessment instruments, and statistical analysis). We conclude our discussion by highlighting the importance of publishing an unbiased and

full report on an EBI's outcomes and not cherry picking or casting aside negative ones. This will help avoid replication failure and unfairly discourage others from using an evidence-based program that might fill an agency's needs.

Historical Antecedents of SFP

The SFP is a unique family-based prevention intervention that combines parenting, youth, and family skills training to reduce adolescent behavioral health disorders including delinquency and the use of alcohol and drugs. The multicomponent program is highly interactive and involves 1-hr sessions earmarked for parents to improve their parenting skills and separately for youth to provide skills that reduce vulnerability to drug use and other behavioral problems (for a review of the logic model and program active ingredients, see Kumpfer, Magalhães, Whiteside, & Xie, 2016). At the conclusion of the separate training sessions, parents and youth come together for another hour of nondirective play to practice and rehearse newly acquired skills, view videotapes exemplifying positive behavior, structure opportunities including role-playing to achieve family harmony, and receive positive feedback from implementation staff.[2]

Briefly, the program blends family systems theory (Forehand & McMahon, 1981; Guerney, Coufal, & Vogelson, 1981) and the social ecology domain model of risk and resilience (Kumpfer & Turner, 1990–1991) to construe youth drug use as part of a "family affair," contextually bound by family dynamics, peer influences, and the social ecology of the home. Using techniques drawn from therapeutic traditions (e.g., Bowen, 1991), social learning theory (e.g., Bandura, 1977), and clinical coaching and skills training methods (e.g., Patterson, 1982), parents are taught effective parenting strategies including how to communicate with their child, setting boundaries and limits (controls and restrictions), appropriately reward their child in a nonpunitive environment, and different ways to bond with their children and increase family cohesion. Following program exposure parents should be better teachers, more empathic, better listeners, and more understanding of their child's world. Children receive training in social and personal competency skills that will help them refuse drug offers and improve their social–emotional regulation, problem-solving, and effective communication. The goal for children is to increase opportunities for identification and bonding with "positive peers, adults, authority figures, and role models" (Kumpfer et al., 2016, p. 70) through greater self-regulation, better conflict and stress management, and learning ways to handle peer

pressure. To be clear, SFP is not a ""one-size-fits-all"" intervention, rather the program has different age versions complementing developmental periods from childhood through high school (SFP 0–3 years, SFP 3–5, SFP 6–11, SFP 12–16, and SFP 7–17) and is adaptive to different child risk levels.

Early Evidence

DeMarsh and Kumpfer (1985) and Kumpfer and DeMarsh (1985) reported on the first trial conducted with a 14-session version of SFP with children aged 6–11. The 4-year trial, funded in 1982 by the National Institute of Drug Abuse, used a four-condition dismantling design randomly assigning chemically dependent parents[3] in treatment and their children either to receive the full SFP, the parent training only, parent training plus a child's skills component, or no additional treatment. The Utah state substance abuse agency subcontracted recruitment to drug treatment agencies who relied on drug counselors to obtain family participation. The outcome evaluation focused on parents' discipline and punishment practices, parent–child communication, and family environment (i.e., harmony) and included a wide range of child behaviors (internalizing and externalizing, delinquency, competence, peer relations, and parent bonding). Parents and children assigned to the full condition offering skills training to both parent and child fared better compared to the remaining three conditions. These improvements included fewer problems reported by parents handling their child with greater awareness of child management strategies. Parents also reported their children were more manageable, showed improvements around the home with fewer behavioral problems compared to their same age peers. Consistent with a reasoned action approach, children reported fewer intentions to smoke and drink, which are important intermediate measures that presage behavior.

Since its initiation, SFP has been tested in 12 RCTs—six conducted with independent research teams (Brody et al., 2006; Brook, McDonald, & Yan, 2012; Coatsworth et al., 2014; Gottfredson et al., 2006; Maguin et al., 2004; Puffer, Annan, Sim, Salhi, & Betancourt, 2017) all producing favorable intervention effects by reducing risk and increasing protective factors that are etiologically linked with alcohol, tobacco, and drug use. The program also improves mental health outcomes, increases personal resilience, reduces delinquency, violence, and aggression and has positive effects on academic performance by reducing school behavior problems including early dropout (Kumpfer, Xie, & Hu, 2011). Importantly, SFP has been shown to work well with all types of families—not just those that are

considered "high risk" including rural (Kumpfer, Alvarado, Tait, & Turner, 2002; Marek, Brock, & Sullivan, 2006) and urban settings (Aktan, Kumpfer, & Turner, 1996) and with different age groups (Kumpfer, Greene, Allen, & Miceli, 2010).

Subsequently, a shorter seven-session "*universal*" version with four booster sessions was developed for youth aged 10–14. This downsized version was developed as part of a university–community collaborative partnership implementing SFP with rural families from Iowa. The modification was instituted for essentially two reasons: First, families recruited for the study found it difficult to travel long distances and attend 14 weeks of classes, increasing attrition (only 20% of the total families recruited participated in SFP 10–14), and second, the Iowa SFP recruited whole 6th grade classrooms through schools to attend the SFP family classes. As a result, the universal program required programmatic changes to reflect the lower risk levels of these families (Kumpfer, Molgaard & Spoth, 1996; Molgaard, Spoth, & Redmond, 2000).[4] Details on the revised SFP 10–14 curriculum can be found at the Iowa State University Extension (http://www.extension.iastate.edu/sfp/).

So Why the Crises of Replication Failure?

As previously mentioned, the problem of replication failure has become a topic of concern not only for psychology in general but also more specifically for prevention science (e.g., Elliott & Mihalic, 2004; Valentine et al., 2011). Replication is both the bane of a scientist's existence and yet the very foundation on which science rests. Findings in one laboratory must be objectively verified in an independent trial, lest subjectivity, and the thought of "tinkering" (or selective reporting) enter the discussion. For prevention science, the discussion over replication has fueled debate over the supposed efficacy of studies when they leave the laboratory and hit the streets. There are many reasons for replication failure, and these make up the balance of this discussion.

In a recent article, Gorman (2017) suggested that regression to the mean, flexible data analysis, and selective data reporting should be added to the list of reasons for replication failure. These contributory factors are germane to all of science not only prevention science (Goodman, Fanelli, & Ioannidis, 2016) and are rooted in philosophical discussion over the process of falsification and the need for corroboration exemplified by Popper (1963). Gorman's claims of replication failure are specific to five independently conducted RCTs testing SFP 10–14. Four of these trials were

conducted in Europe and one in the United States, which as he argued did not show evidence of favorable program outcomes and provide further evidence of what he termed a *decline effect*.[5] That is, after the initial hoopla, there is a preponderance of evidence supporting failed replications. The four European studies with null effects were conducted in Germany (Baldus et al., 2016; Bröning et al., 2017), Sweden (Skärstrand, Larsson, & Andréasson, 2008; Skärstrand, Sundell, & Andréasson, 2014), Poland (e.g., Foxcroft, Callen, Davies, & Okulicz-Kozaryn, 2017; Okulicz-Kozaryn & Foxcroft, 2012), and Wales (Segrott et al., 2017). The U.S. study was conducted in the Midwest (Riesch et al., 2012). Gorman then dug deep into the history of trials supporting the efficacy of SFP 10–14 including a long list of RCTs conducted by the Iowa team. These RCTs involved the seven-session SFP 10–14 in comparison to the five-session *Preparing for the Drug-Free Years* (Redmond, Spoth, Shin, & Lepper, 1999; Spoth & Redmond, 2002; Spoth, Redmond, & Shin, 2001; Spoth, Redmond, Shin, & Azevedo, 2004), in conjunction with Botvin's *Life Skills Training* program with and without SPF 10–14 as a comparison group (Spoth, Randall, Trudeau, Shin, & Redmond, 2008; Spoth, Randall, Shin, & Redmond, 2005; Spoth, Redmond, Trudeau, & Shin, 2002), or a combination of the three programs (Spoth, Trudeau et al., 2008). All of the studies included longitudinal follow-up, some extending from late childhood through adolescence (Spoth et al., 2001; Spoth, Randall, et al., 2008) and others extending to young adulthood (e.g., Spoth, Trudeau, Guyll, Shin, & Redmond, 2009; Spoth, Randall, et al., 2008).

Reasons for Replication Failure

Notably, the international studies and the Midwest trial were free of conflict of interest concerns over developers evaluating their own program (Gorman, 2005), provided evidence of high quality implementation, and all adhered to the published standards of evidence (Flay et al., 2005). Given these accolades, a pressing question is why the replication failure and what contributes to what Gorman termed a decline effect? To begin with, one of the big leaps that programs must make is moving from efficacy trials conducted in tightly controlled laboratory settings to effectiveness trials, the latter conducted in community agencies where strict protocol replication is rendered more difficult because of a variety of competing factors and circumstances (e.g., Glasgow, Lichtenstein, & Marcus, 2003). These can include staffing, recruitment (consenting and enrolling), attrition, and financial problems as well as competing organizational interests including taxing

staff work schedules, lack of organizational buy-in, poor implementer effort or poor training and supervision, lack of stakeholder engagement, or a program "champion," all of which can diminish program enthusiasm. Below, we outline the six most pressing concerns that affect the delivery and implementation of EBIs. Included in this discussion are strategies that agencies can utilize to overcome these hurdles when implementing family EBIs like SFP.

Matching Program Type to the Target Audience and Implementing With Fidelity

Replication failures can occur because an agency implements an EBI without considering whether the risk levels in their population match those in the original research trials. This "lack of fit" can increase the chances for failed results. It is well known that EBIs are primarily effective when they are delivered to populations similar to those participating in the original RCT. In many instances, agencies select a program that addresses their workplace demands. This can result in their choosing the shortest program given personnel and resource constraints. The end result is the program selected does not match the required dosage to address relevant family needs and risks. The seven-session SFP 10–14 is slated to work well in universal settings with low-risk families. In the case of the Iowa program, even with the downsizing from 14 to 7 sessions there is sufficient attention given to risk and protective factors, and enough time is spent rehearsing newly acquired skills to have an indelible impact on family dynamics. A different setting, as many of the replication studies encountered, with higher risk families may require greater dosage obtained through additional sessions. The additional sessions provide more opportunities for learning program content as well as practice opportunities to rehearse skills and receive feedback.

Many have observed that higher risk families participating in family interventions will experience larger positive changes because they have more room for improvement. Therefore, qualification of "risk" status plays a major role in determining the effectiveness of SFP, as it would any selected or indicated program. In some cases, SFP replications recruited low-income families assuming this rendered the population at risk (Bröning et al., 2017). This can be problematic because efforts to recruit from high-risk or ethnic neighborhoods does not necessarily ensure that families are deficient in family skills. In other words, by itself, poverty or racial minority status does not necessarily equate with family risk.[6] This is particularly true

when low-income families access psychological resources that buffer risk (e.g., Burchinal, Follmer, & Bryant, 1996; Markstrom, Marshall, & Tryon, 2000). Furthermore, structural aspects of families such as single parenting may be "markers' of risk but not equated with risk in and of itself (Rutter, 2006). In the long run, risk is context-specific and this has to be factored into the choice and selection of a program. Overall, researchers working with family-based interventions obtained better results with distressed than with not distressed, families (e.g., Dishion, Nelson, & Kavanagh, 2003). The European replications used the seven-session SFP 10–14, a choice guided partly by the Cochrane meta-analysis findings showing excellent results with school-based alcohol prevention using the seven-session Iowa SFP 10–14 (Foxcroft, Ireland, Lister-Sharp, Lowe, & Breen, 2003). However, despite such success the program may not have suited the population's risk level. For instance, Riesch et al. (2012) pointed out that high-functioning families may have "limited potential gains" (p. 367) from a brief intervention like SFP 10–14 and, furthermore, the very low levels of substance use by youth in this age-group may render the content more abstract than practical among parents who don't view their children as deviant or needing extra supervision.

Administering the Right "Dosage" of the EBI

Reduced dosage. Integrity of the treatment, or what is termed dosage, can heavily influence program outcomes.[7] To us, any modifications from the intended dose will needlessly dilute program effects. There is no formula for calculating the appropriate dose based on target sample characteristics, and as a result the program should be implemented in full form to achieve maximal effects. However, it is fairly common in community agency or school settings to reduce the dosage of SFP to match local risk levels in what has become termed "off label" use. However, dosage analyses frequently show that curtailing sessions or eliminating booster sessions reduces effect sizes (Riesch et al., 2012). Indeed, Riesch and colleagues opined that effects in their study may have been larger had they chosen the 14-session SFP 6–11 program, which is appropriately targeted to higher risk families; a position driven partly by the observation that they reported better program outcomes with a high exposure group (attending >5 sessions of 7).

Program modifications come in various shapes and sizes and can include selectively implementing lessons, changing exercises, eliminating or, in some cases, adding new lessons. The Swedish replication eliminated skills training but added more drug prevention lessons (Skärstrand et al., 2008,

2014). Another example is the Hawaii SFP 6–11, where staff added 10 extra preliminary sessions on Hawaiian cultural values while eliminating regular SFP skills sessions (Kameoke, 1996). This resulted in high attrition and poor results until staff restored the original 14-session program creatively infused with Hawaiian values. A key factor in understanding the role of dose in program outcomes is to frame effects as the result of an unambiguous treatment comparison. In other words, we should not compare SFP as a treatment to a modified program with fewer sessions. We should compare SFP administered in its entirety to a no-contact control condition. It is generally not recommended to compare treatment conditions that vary dosage levels (based on quantitative cut points).[8] SFP should be delivered in its entirety with no modifications to the dosage of the program as this maintains curriculum integrity and keeps intact the active ingredients as they were designed (Small, Cooney, & O'Connor, 2009). This strict program adherence will also contribute to an unambiguous interpretation of the effects that result from intervention exposure.

Properly Training Implementers and Building "Enthusiasm"

Type II errors can also arise from procedural variation in implementation. This is a major concern when transitioning from efficacy to effectiveness trials because there is so much greater control exercised in efficacy trials. This control extends to implementation, which is much easier to regulate in a controlled trial setting than in real-world conditions using community agency personnel. Although there are a host of organizational challenges that can influence implementation, three in particular that are common issues facing SFP implementation include misconstruing quality for fidelity, heterogeneity in implementer experience, and poor quality of implementer training. We briefly discuss each of these considerations below.

Quality is not the same as fidelity. Most SFP replications report a high degree of fidelity to the model. However, implementing with "fidelity" is not necessarily implementing with quality and enthusiasm. The Washington, DC, SFP replication is a good example of this (Gottfredson et al., 2006). Hired community workers documented fidelity through site visits with videotapes of live sessions plus fidelity checklists that monitored program adherence (assessing whether the implementers followed the training manuals). However, this process evaluation suggested lackluster implementation with little enthusiasm or quality, which would diminish favorable program outcomes. Enthusiastic and competent implementers who are "true

believers" in the effectiveness of SFP provide genuine feedback to participants, encouraging them to engage in activities, providing examples to build foundational skills, and ensuring that participants are eager to return for future sessions. In separate analyses, we correlated SFP outcomes with facilitator characteristics and found that staff who are very enthusiastic about the program and committed to helping the families improve obtain the best outcomes (Kumpfer et al., 2017; Orte, Ballester, Torelló, de Vicente, & Mascaró, 2017).

Implementers not experienced. In the real world, implementer experience can vary dramatically, but most agency staff are quite competent working with families that utilize their agency services. When SFP is implemented with family service agencies, the results match those found in tightly controlled RCTs. Two replications reinforce this claim including a 5-year study conducted in 75 community agencies in New Jersey and involving all four age versions of SFP (Kumpfer, Greene, Allen, & Miceli, 2010) and the 10-year multisite study of SFP 12–16 still underway in Ireland (Kumpfer, Xie, & O'Driscoll, 2012). The Irish study has obtained consistent positive results for both high-risk girls and boys using an agency collaborative model with staffing and family recruitment shared by probation services, alcohol and drug treatment and prevention, mental health, family services, and the local police.

Poor quality of implementer training. Staff training is critical to the success of any family-based intervention, and when suboptimal can invariably undermine program fidelity and weaken program outcomes. This is particularly crucial when working with cultural adaptations or implementing programs in foreign settings. We always translate all training materials into the language of participants and modify the curriculum graphics to appeal to the target population. When we implement SFP in foreign countries, we use trainers fluent in the language and familiar with local customs and mores, educational practices, and family needs. This was the case for an independent, large-scale RCT of SFP 6–11 with Burmese refugees in Thailand that relied heavily on local trainers (Puffer et al., 2017) and was evaluated by an independent team at Duke University using assessment tools modified in language and content, with outstanding results. SFP uses a "train-the-implementers" approach and we used role-playing with parents and teens to address cultural issues that might crop up during real-time training with parents and youth. The training focuses on teaching implementers how to build rapport, confidence, and trust with the families. We also ensure there

is optimal organizational "buy-in" to support the training and that agency leaders are aware of the commitment (monitoring and evaluation) needed to implement with fidelity.

To summarize, it is possible to achieve high levels of enthusiasm when training program implementers. This can be achieved through education, working with the teams to ensure cultural adaptations reflect the community needs and sensitivities, and also by providing technical assistance and feedback not only at the beginning of training but throughout as teams encounter problems and need direction. Training manuals should explicitly address commonly encountered problems (i.e., family reticence) but provide a formal basis for "standardization" of the program, so that novelty can be incorporated without diluting the objectives. Additional strategies to build enthusiasm include soliciting administrative support, so that workers feel their back is covered, commitment of resources, and ensuring the program is perceived as credible before implementation. Overall, enthusiasm is best recognized as motivation to help the clients but packaged in a way that allows the strengths of the program to speak volumes on its own.

Maintaining Program Integrity

Lack of fidelity to core elements of the model can also affect outcomes. In addition to dosage considerations, research shows that too much modification can denude a program of its "active ingredients" and weaken effects. This is precisely what occurred with the Swedish SFP 10–14 RCT, perhaps contributing to what is termed the decline effect. The high costs associated with remaking the DVD for Swedish families left sparse funding available for implementation. As a result, program modifications were made, which affected dosage, delivery of core competencies, and ultimately disrupted the program's *deep structure* (Resnicow, Soler, Braithwaite, Ahluwalia, & Butler, 2000). In the Swedish case, SFP was no longer truly a family skills intervention involving participation of whole families, where parents and youth enjoyed a meal together and practiced newly acquired skills. To accommodate their lean financial resources, the Swedish program was severely curtailed by eliminating the joint parent–youth practice sessions. Instead, parents met separately at night and school teachers conducted the youth portion with their whole classrooms of 25–30 youth. The absence of prescribed group activities coupled with poor classroom management most likely diminished program effects. Also, regular lessons were dropped and extra substance use education lessons were added. All of these factors can

attenuate program effects as both fidelity and dosage are compromised (Segrott et al., 2014).

Indeed, Ferrer-Wreder, Adamson, Kumpfer, and Eichas (2012) and Segrott et al. (2014) questioned the value of the Swedish SFP 10–14 intervention as a direct replication or even a family EBI at all given the substantial modifications to both content and implementation. By dropping various crucial elements of the intervention, the Swedish team inadvertently conducted a componential analysis selectively culling key ingredients. The result suggests that the critical core element of SFP requires attendance of the whole family (other siblings and caregivers) together and that program outcomes are inextricably tied to core features of the program including provision of meals together with practice time allotted for rehearsing new skills, parent–youth weekly practice sessions with facilitator feedback, and participation in family homework assignments, all of which were absent in the Swedish replication.

Culturally Tailoring the Program

Cultural adaptation is an essential part of the fidelity discussion. Family-based programs delivered to different cultural populations require some form of adaptation (e.g., Castro, Barrera, & Martinez, 2004; Kumpfer, Pinyuchon, de Melo & Whiteside, 2008). Here again, research shows that cultural adaptation with family-based programs will increase enrollment and program completion (e.g., Kumpfer et al., 2017). We are aware that there is a fine line between instituting needed cultural adaptation without necessarily instituting complete program modification. Nonetheless, whenever SFP is culturally adapted at the surface level (Resnicow et al., 2000) including enhancements to incorporate the local language, myths, relevant exercises, games, songs, and rewards, SFP has better recruitment, less attrition, and better outcomes (Kumpfer, Alvarado, Smith, & Bellamy, 2002). This is also true of other family EBIs such as multisystemic family therapy, which also found better recruitment and less attrition when the core intervention strategies were culturally adapted (Parra-Cardona et al., 2016). In the case of *Familias Unidas*, an integrative program specifically constructed for Hispanic immigrant families and targeting youth substance use (Pantin et al., 2003), specific modules are introduced that address acculturative stress and parent–child tensions associated with immigration (i.e., reducing barriers associated with moving to a majority culture).

There is now evidence that culturally adapted SFP versions applied in non-European cultures have achieved high participation rates and excellent

outcomes (Puffer et al., 2017). Added to this, quasi-experimental studies conducted recently in Spain (Orte et al., 2017), Italy (Oretega, Giannotta, Latina, & Ciairano, 2012), the Netherlands (Onrust & Bool, 2006), and Ireland (Kumpfer et al., 2012) produced favorable program outcomes with the SFP 6–11 and SFP 12–16 programs including large effect sizes. Unfortunately, since these programs do not meet the gold standard of RCTs, researchers conducting meta-analysis will invariably overlook incorporating these favorable SFP findings.[9]

Invariably, there are also examples of programs that fail to institute cultural adaptations and produce less than stellar outcomes. For instance, Olds et al. (1997, 2004) successfully conducted three RCTs in the United States of the *Nurse Family Partnership*, a program targeting low-income primiparous mothers to reduce dysfunctional caregiving. However, a U.K. replication study failed to obtain the same results (Robling et al., 2016), which the authors attribute to lack of cultural adaptation and poor fit to the risk levels of the target population. Failure to culturally adapt a program can also promote resistance on the part of the staff. The Washington, DC, SFP replication (Gottfredson et al., 2006) provides an example where an aborted cultural adaptation resulted in poor outcomes. In this case, the implementation staff were told very early in the process, prior to the first year of implementation, they could begin preparing a cultural adaption of SFP 6–11 targeting African American families. However, for various reasons,[10] the cultural adaptation was aborted, which lessened staff enthusiasm for the project and diminished their quality of delivery, with the end result of less favorable outcomes.

Cultural adaptation runs across several of the themes we have already discussed including staff training, supervision, process evaluation, and program materials (language translation). Each of these factors into the success of the program but also can weaken attempts at cultural adaptation if not done correctly using rigorous methods. For example, certain issues relevant to the core competencies of SFP may not translate directly into another language. Also, staff may be reluctant to discuss certain sensitive issues with families given underlying differences in cultural mores. Social contexts and family dynamics can also vary between cultures, making it prudent to flesh these issues out in preliminary field work prior to implementation or through protracted discussion with staff (e.g., Akin et al., 2016). At the staff level, many cultures will lack experience conducting process evaluations using formal instruments to gauge fidelity. All of this needs to be considered before implementation, otherwise it can create roadblocks and hinder obtaining successful program outcomes. In the long

run, cultural adaptation is not something "stock" that comes off the shelf but rather involves a lengthy iterative process of program modifications that involves extensive checks and balances to ensure the program does not sacrifice fidelity for "fit" (Barrera & Castro, 2006).

Choose the Right Outcome Evaluation Methods

Evaluation practices for family-based programs implemented in community agency settings are much like an onion; evaluation teams have to learn to peel away the layers, using different assessment and data collection strategies and measurement tools to discover the different factors that influence program outcomes. In certain situations, as evidenced by the Washington, DC, SFP replication, extensive in-person interviews can reveal subtle influences on program delivery that affect outcomes not evidenced by traditional statistical analysis. Such efforts can involve the parents, their children, or implementation staff and may require a mixed-methods approach that blends qualitative and quantitative assessments strategies to reveal the full gamut of influences on program outcomes. Included in this process is using evaluation methods that are developmentally appropriate for the target audience and culturally appropriate for both the implementation team and the participants.

Selecting outcome measures not developmentally appropriate. It is likely the five international SFP 10–14 replications failed to achieve statistically significant differences in drug use because of low base rates characteristic of low-risk 12- to 14-year-olds. For instance, many low-risk children do not use drugs producing extremely skewed frequency distributions. Although statistical modeling approaches can be applied to correct for skewness (e.g., Olsen & Schafer, 2001), they are not a panacea. In many cases, using theoretically consonant intermediate measures (i.e., intentions to use) provides an alternative to model program effects. The 14-week SFP replications are more successful in achieving larger effect sizes, probably because they target higher risk 12- to16-year-olds or children in drug involved families. Also, the low-risk parents attending SFP 10–14 were highly unlikely to demonstrate the poor parenting or family skills that are the focus of SFP, making it harder to show improvements. Notwithstanding, low-risk participants can still benefit a great deal from skills training activities that are core features of SFP (i.e., parent–child communication, setting boundaries, and family organization).

Using clinical diagnostic instruments that are not change sensitive. Program evaluation instruments should be sensitive to change using at least 5-point Likert-type scales. Most clinical diagnostic instruments like the *Child Behavior Checklist* and *Strengths and Difficulties Scale* only have a 3-point scale designed as clinical diagnostic instruments and not for evaluation. They are intended to provide prevalence data but lack the subtleties required for monitoring true behavior change. It is worth pointing out that the modified response formats were used in the European SFP 10–14 replications. Changing response formats can truncate variances, and with changes in dispersion and first-order moments render findings nonsignificant.

Use longitudinal repeated measures control group designs. Longitudinal follow-up is required to discern whether a program has sustained effects over the long haul and to account for confounding by developmental maturation (e.g., Collins, 2006). Although SFP can be delivered in the elementary school years, typically, youth encounter peer pressure to use drugs beginning in middle school and this rapidly increases through high school. Higher risk youth may encounter these pressures earlier, given their peer group may express deviant behaviors at an earlier age. Several replications have shown favorable effects when youth were followed through middle school. For instance, Bröning et al. (2017) reported 11 of the 18 positive outcomes in subgroup analyses comparing high- and low-risk youth in the German trial of SFP 10–14 with 2 years of follow-ups to age 14 years (see also Baldus et al., 2016). Abstinence outcomes for tobacco, alcohol, and cannabis had small effect sizes (.10 to .16); however, they needed a larger sample size of 785 versus the 135 that participated to reach statistical significance. Low power is characteristic of many SFP 10–14 replications leading to erroneous conclusions the program did not work.

Using regular pretest and posttest instruments. Years of experience assessing SFP outcomes have shown that regular pretests underestimate family risks at program entry. This diminishes the amount of positive change and effect sizes by posttest or subsequent follow-ups. We mainly use a standard pretest as well as a retrospective pre- and posttest conducted after program graduation. We found that even parents who had lost their children to foster care would rate themselves at pretest as wonderful parents with well-behaved children. Following exposure to self-monitoring assignments and parenting skills activities, parents are more aware of their deficiencies and rate themselves lower. Brook, Akin, Lloyd, Bhattarai, and McDonald (2016)

provided evidence that the retrospective pretest is more accurate and matched the implementers' ratings of the families. The retrospective pretest–posttest design can provide more "veridical" assessments of self-behavior, owing to giving participants realistic anchors and avoiding response shift bias (Chang & Little, 2018).

Provide an Accurate and Complete Outcome Evaluation Report

The value of statistical significance versus clinical significance or effect size to determine effectiveness. Tradition suggests that the benchmark for a valid intervention effect is statistical significance set by a p value below .05 rather than clinical significance measured by effect size or how much clients changed. Replication failures often arise because of low power or small sample sizes that prevent a statistical comparison from achieving statistical significance (Maxwell et al., 2015). Relatively, large sample sizes observed in the original RCTs conducted by Spoth and colleagues were able to produce statistical significance with small effect sizes. However, these results are not replicable with smaller studies that may involve a handful of agencies that bundle their intervention efforts together. Hence, Tryon (2016) suggests that, as a general rule, we should avoid relying on single studies and especially avoid comparisons to generously funded university-based RCTs with large sample sizes.

A central factor when considering EBIs should consider effect sizes, which should receive equal if not greater attention than statistical significance.[11] This consideration should extend to outcomes as well as putative mediators. Furthermore, a significance test can obtain p values below the nominal .05 even if effect sizes are small if the sample is sufficiently large. To avoid criticism, evaluators should publish effects sizes bounded by confidence intervals for the full gamut of outcomes, which is more useful to clinicians determining which interventions work best and under what conditions (e.g., Simmons, Nelson, & Simonsohn, 2011). This is the preferred strategy that Kumpfer, Magalhães, Whiteside, and Xie (2016) followed when they published all 18 parent, family and child outcomes, and their effect sizes (which are medium to large size) for each 14- or 10-week SFP study.

Strategically, evaluation analyses should include a careful examination whether the program has sizable effects on parenting skills, family relations, and child or youth outcomes other than substance abuse. There is considerable evidence showing that SFP has favorable program effects on depression, overt and covert aggression, delinquency, and school performance

(Kumpfer et al., 2017). These developmental outcomes are just as important as substance use because research shows they have a direct effect on substance abuse. This latter view is consistent with problem behavior theory (Jessor & Jessor, 1977), which posits there is a constellation of negative developmental outcomes sharing common etiological pathways. It is also a mainstay in developmental psychopathology, which uses the concept of equifinality to support how multiple pathways can lead to a single outcome and, in some cases, a single pathway, referenced as multifinality, can lead to divergent outcomes (Cicchetti & Rogosch, 1996). In this case, correcting family dynamics as a singular focus can produce multiple favorable outcomes.

Why Real World Replications Can Fail to Replicate Original RCT Results

There are a number of other reasons for replication failure that still need to be addressed. Contamination of the no-treatment control group is one of several threats to internal validity and can happen for a variety of reasons. Contamination can occur because trainers within a single family service agency encounter and work with both intervention and control cases on a daily basis. In the Washington, DC, SFP 6–11 replication, the implementation team sometimes applied SFP or other clinical techniques to assist minimal contact control families leading to improvements in control families. This causes diffusion of treatment and violates the stable unit treatment value assumption required to maintain causal inferences (Rubin, 2005). Some solutions to prevent contamination are increase the number of trainers, which is often cost prohibitive; increase the number of family service agencies and clinics to avoid diffusion of treatment; or have minimal face-to-face contact with the control group until time to take the posttest.

Including intention-to-treat families in the analyses can be misleading. Several SFP replication studies have relied on intention-to-treat analyses. In this scenario, families are included in the data analysis based on their original experimental assignment irrespective of their continued participation throughout the duration of the study.[12] This methodology is traditionally used with clinical trials and public health initiatives to capture information from missing subjects irrespective of their exposure levels (e.g., Heritier, Gebski, & Keech, 2003). Unfortunately, this approach to data analysis will include families (parents and youth) regardless of whether they attended

one or two sessions or even none at all (e.g., Dishion, Kavanagh, Schneiger, Nelson, & Kaufman, 2002). Since dosage is apparently a major factor in program success, this reduces the outcome effect sizes of those that actually attended most or all sessions. Program evaluators need to consider the influence of missing sessions and find ways to differentiate this phenomenon from lack of exposure that can arise from poor fidelity (the program is delivered but poorly). Then, they need to figure ways to properly evaluate the program for participants that were present and received a majority of the treatment (i.e., high fidelity analyses). There is now more work ensuing that uses inverse probability weighting and propensity scoring methods to adjust postrandomization for session attendance or dropout status and that bears on the issue of *exposure* (e.g., Little & Yau, 1998; Tein et al., 2018). Following these, few recommendations will inform the public whether the program achieves its objectives when delivered efficiently and with fidelity.

Inaccurate Outcome Reporting

Selective data reporting. Utilization of *flexible data analysis* can also contribute to Type II errors. By this, Simmons, Nelson, and Simonsohn (2011) mean that dicing up analyses to accommodate elimination of certain subjects, effectively examining minimal dosage requirements, conducting subgroup analyses using high fidelity participants, or creative manipulation of outcomes (i.e., dichotomization; MacCallum, Zhang, Preacher, & Rucker, 2002), all of which can inadvertently bias statistical findings and increase the incidence of false positive rates. Simmons et al. suggest that there is no malicious intent here but "ambiguity" that comes along with making decisions how to approach data analysis. Their simulations show convincingly that even increasing the sample size by adding 10 observations or controlling for a covariate and its interaction can appreciably increase the false-positive rate. All of this leads to the conclusion that there is a need for greater transparency in data decision making. This arises because of the "researcher degrees of freedom" (p. 1359) or stated differently, many program developers are heavily invested in finding out whether their intervention produces favorable outcomes (i.e., lowered drug prevalence rates) no matter the framework for analyzing the data. In other words, most researchers believe that no matter the approach, or the cost, it is worth finding out if their program works in *some way or another*.

Consistent with these considerations, we recommend avoiding selective analyses unless they are consistent with theory, hypothesis driven, and represent promising avenues of inquiry. We also hold that there are analysis

strategies that require explicit rationales. For instance, many researchers examine gender subgroup analysis without providing explanations for why a program should work differently for boys and girls. Programmatically speaking, this requires some attempt at using gender socialization to argue unique pathways or differential program outcomes. Likewise, race/ethnic group is often used to calibrate program effects without providing explanation for the observed differences. Why would Black or Hispanic youth react differently to the program content or implementation? This requires more careful thinking that has strong theoretical roots and considers race-specific contextual factors. This line of reasoning extends to using high versus low fidelity groups in analyses to determine moderation of program effects. It should be clear that parents/children receiving high dosages may fare better, but there is no hard and fast rule on what constitutes a sufficient dose. More research is needed to determine whether there are critical cut points that qualify necessary and sufficient dosages. Likewise, tabling all of the designated outcomes for both parents and children presents the best case scenario rather than selectively reporting outcomes only achieving significance. In addition, flexibility pertains to conducting multiple post hoc tests and using one- versus two-tailed tests. The problem of multiplicity in analyses is rampant in science and has come under criticism before (e.g., Goodman et al., 2016). To avoid p-hacking, the norm should be a well-designed set of analyses to address the stated hypotheses with appropriate one-tailed tests, since we don't expect the program to have an iatrogenic effect and increase drug use (e.g., Head, Holman, Lanfear, Kahn, & Jennions, 2015).

Failure to report all SFP outcomes. Claims of replication failure are tied to the assertion that SFP 10–14 replications were unable to show reductions in rates of substance use. As we already stated, this strategy emphasizes a focus on substance use outcomes, which are unlikely to show marked change in low-risk students reporting nominal amounts of drug use. We have also stated in numerous places that reductions in substance use is not the sole goal of SFP, which also focuses on mental health and behavioral disorders, child maltreatment, school performance, and other related developmental problems that interfere with normal functioning (Kumpfer et al., 2016). The action theory for SFP posits the program works *generatively* through putative mediators (i.e., parenting, family, and youth skills) that are formative in protecting youth against a wide range of negative developmental outcomes. In keeping with this view, it pays to examine short-term measurable goals on putative mediators, many of which are precursors to

youth substance use. This was the case in the five cited replications (Bröning et al., 2017) and evidenced in other family EBIs (e.g., Van Ryzin, Kumpfer, Fosco, & Greenberg, 2016). This emphasis is partly guided by a national mandate and public health demands (U.S. Department of Health and Human Services, 2016).

Many of the anticipated changed in mediators also contribute to positive developmental outcomes requiring that we also inspect changes in these behaviors (e.g., school performance) to note the different directions that improved family functioning can take. This latter position is consistent with developmental cascade models that are today's norm in both etiology (e.g., Eiden et al., 2016) and prevention (Patterson, Forgatch, & DeGarmo, 2010). There are several prime examples of the cascading effect of SFP on important aspects of positive youth adaptation. For example, the Safe African American Families program, a modified version of SFP 10–14 (Kogan et al., 2016) found 50% reductions in diagnosed depression and anxiety, substance abuse, criminality, and HIV status at 10-year follow-up. These favorable outcomes were observed in the 30% of the participating youth who had genetic risks for various disorders determined by a saliva test (Brody et al., 2012). Moreover, a separate independent evaluation of SFP 3–5 and 6–11 as part of a multistate trial showed favorable program outcomes with reduced child maltreatment and days remaining in foster care cut by half (Brook et al., 2012). A cost recovery study found millions of dollars were saved with SFP (Johnson-Motoyama, Brook, Yan, & McDonald, 2013). Certainly, we should not dismiss or discount these promising findings. Additionally, there are studies that augmented SFP with additional mindfulness training and reported favorable outcomes (Coatsworth et al., 2014). All in all, and given the heavy modifications to the core SFP competencies, these studies may not represent "direct" replications; however, they do provide additional evidence of the basic effectiveness of the unique SFP family skills training model.

How Can We Know What Really Works?

One solution is to use small clinical RCTs. Since large-scale RCTs are very expensive, another solution would be repeated small scale RCTs conducted in clinical settings with real clients. This approach was used for Triple P (Sanders, Baker, & Turner, 2012), which is listed on several EBI websites. The authors used a short-term waiting list experimental design. Dynamic wait list, rolling recruitment or a stepped wedge design are approaches that more family EBIs should consider. This is the approach we took with SFP

7–17 using agencies in a three-state study including NY, NC, and UT. Other possible solutions to increasing knowledge of what works in reality is to broaden the definition of effectiveness. Reviewers or raters preparing listings on websites of EBIs need to consider more than just medical model RCTs as proof of effectiveness. What should matter more are not Phase III clinical efficacy trials, but multiple Phase IV effectiveness trials in the field to determine whether the EBI works with diverse clients in diverse settings. Thereafter, the next phase involves Phase V dissemination trials when going to scale with large numbers of clients. As a general rule, we should all be more inclined to consider the *total weight of the evidence* for an intervention obtained from multiple studies (Goodman et al., 2016), and hopefully ones that are conducted by independent research teams.

Should Replication Failure Worry Us?

Should we be worried by the reported failures? The answer is both "yes" and "no." We should certainly attempt to glean more information regarding why failure occurs overall, and so that other agencies don't make the same mistakes. While the five independent replications failed to completely replicate the SFP 10–14 program effects on drug use, SFP has a long history of positive results. For the reasons outlined in this article, we believe these failures are symptomatic of poor implementation and do not undermine the program's integrity or its capabilities. As Strobe and Strack (2014) pointed out, "Even multiple failures to replicate an established program finding would not result in a rejection of the original hypothesis, if there are also multiple studies that supported that hypothesis" (p. 64). Since there are multiple studies supporting SFP 10–14 effectiveness, there is sufficient evidence supporting favorable SFP outcomes obtained from RCTs with the 14-session SFP (Kumpfer et al., 2002; Puffer et al., 2017), evaluations relying on propensity matching techniques (Brook et al., 2016) and numerous quasi-experimental, large sample field studies (Kumpfer et al., 2010, 2012).

Reichardt (2011) points out that in terms of evaluation science, there is a tremendous difference between asking "what is the effect of a given cause?" as opposed to "what is the cause of a given effect." Program evaluation attends only to the first question, whereas the second question, while also quite compelling, is reserved for asking what happened during implementation that may have affected the outcomes. This crucial distinction is often not made when evaluating programs but is necessary to find the fine line that divides why some programs work in efficacy trials and then don't

replicate in effectiveness trials. In this article, we have discussed ways to avoid replication failure, protect against the decline effect, and eliminate Type II errors that result in negative findings with a program that has a proven track record and is regarded as "evidence based." In essence, we bridge the chasm between Reichardt's causal statements showing that greater attention should be paid to factors mitigating program efficacy and creating a closer alliance between implementation and prevention science (e.g., Wandersman et al., 2008). In this regard, family-based programs should not be selected based solely on cost or length, as if shorter programs are going to be equally effective as lengthier programs. Dosage and content of a program has to match the risk profiles of the clients. This may go a long way toward avoiding the problems we addressed here, regarding the effectiveness of programs with low-risk populations. The one-size-fits-all approach may not work with family-based interventions, which may require that program content is tailored or carefully customized for different populations.

Family-based programs administered in real-world settings should be implemented with adequate support services; ensuring organizational buy-in; gathering stakeholder support; using well-trained, experienced, and committed implementers who are considered program "champions," implemented with fidelity (not cutting out sessions or adding new untested ones), maintaining proper dosage fitting the target populations' needs, and culturally adapting the program using community-participatory strategies that enhances buy-in. Frequently, financial and market factors often dictate selection of EBIs without concern for length, appropriateness, or the fit of a program to the risk levels of the population. The truth is that the program contents, structure, and delivery mechanisms (including dosage) has to match the risk level of the clients. Hence, applying EBIs developed and tested for universal low-risk populations should not be considered effective for high-risk populations until tested with that population. Program effects that vary based on subgroup status (higher vs. lower risk) are considered compensatory for higher risk groups and leveraging for lower risk groups. Regardless of distinction, evidence of moderation of effects by risk status lessens the ability of a program to be regarded as truly universal (Spoth, Shin, Guyll, Redmond & Azevedo, 2006).

When deliberating these choices we need also consider not only the final outcome measures but also mediators or precursor variables that are targets of the intervention. Examples of mediators should include improvements in parenting skills and family relations and examples of alternative "precursors" could include child or youth outcomes such as depression,

overt and covert aggression, as these are proven antecedents to youth substance abuse and delinquency (Kumpfer et al., 2016). Overall, there are a myriad of factors that impinge on the success of a program when implemented in real-world settings. Only when all of these factors have been considered can we really know the truth about program effectiveness.

Notes

1. These overlap somewhat with the 11 principles of program effectiveness outlined by Small, Cooney, and O'Conner (2009) but depart somewhat based on our experience of implementing family-based and evidence-based interventions.
2. Actual program length is 2.5 hr per session with a coordinated meal for the first half hour. The sessions are led by gender-balanced and ethnically matched trained implementers.
3. The trial was designed as a substance abuse prevention strategy for parents with opiate, narcotic, and polydrug use dependencies. The parents readily recognized they were dysfunctional, spending less time with their child, frequently using negative punishment, and lacking positive parenting and child management skills. The program focused primarily on parenting skills training but includes some drug education taught using didactic methods. By all accounts, this version of the program was "selective"; however, it has since been recast as a universal prevention program, targeting lower risk families with fewer personal, and child management problems.
4. Dr. Kumpfer was the Co-PI on the ISU grant and PI on the original 1982 NIDA grant (R01 #DA02758-01/5), "Prevention Services to Children of Substance Abusing Parents." She worked collaboratively with researchers at Iowa State University to design SFP 10–14; however, she had no direct involvement in reporting outcomes of the Iowa SFP 10–14, which is copyrighted and marketed through Iowa State University.
5. This term was originally used by Reeves and Rhine (1943) as part of their parapsychological research conducted at Duke University.
6. Hill's (1972) classic examination of resilience among inner-city Black families makes this point.
7. In some circles, this is also called adherence and fidelity. Here, we mean only to discuss the integrity of the treatment at a more global level in terms of the amount of exposure the participant receives (i.e., contact hours or sessions and their respective intensity). Fidelity and adherence go beyond this definition to include the way the program is taught, how closely the implementation staff adhere to the training manual, and programmatic modifications made on the spur of the moment (delivery of program content on a session-by-session basis).

8. This would be consistent with a regression-discontinuity design using a predetermined cutoff value. Experimental comparisons would then contrast different "dosage levels" to determine effects.
9. Interestingly, Gorman selectively excluded several randomized control trial that emphasized cultural adaptations or studies of implementation. He also excluded studies without control groups, albeit we learn a great deal from implementation studies because the helps to paint a more vivid picture of real-world concerns that can interfere with successful execution. The end result is that we know a great deal about efficacy but little about effectiveness.
10. A prime reason for abandoning the cultural adaptation was the extra experimental condition would reduce power in a design that already had four conditions. Other factors diminishing program effects may have included contamination of the minimal contact control condition, relatively high staff turnover, poor staff training, and high community disorganization, to name a few, all of which adversely affected program outcomes.
11. There is considerable debate in the psychological sciences regarding the value of null hypothesis testing and the value of effect sizes as opposed to significance testing (e.g., Nickerson, 2000).
12. One factor that contributes to this situation is noncompliance that arises from participants' crossing over between treatments regardless of whether the staff caused this to occur or other reasons like resentful demoralization or compensatory rivalry. Regardless of the origin of noncompliance, motivational factors that differentiate participants represent "selection differences" postrandomization that need to be controlled statistically.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Lawrence M. Scheier http://orcid.org/0000-0003-2254-0123

References

Akin, B. A., Brook, J., Lloyd, M. H., Bhattarai, J., Johnson-Motoyama, M., & Moses, M. (2016). A study in contrasts: Supports and barriers to successful

implementation of two evidence-based parenting interventions in child welfare. *Child Abuse & Neglect*, *57*, 30–40.

Aktan, G., Kumpfer, K. L., & Turner, C. (1996). Effectiveness of a family skills training program for substance abuse prevention with inner city African American families. *International Journal of the Addictions*, *31*, 158–175.

Baldus, C., Thomsen, M., Sack, P. M., Bröning, S., Arnaud, N., Daubmann, A., & Thomasius, R. (2016). Evaluation of a German version of the strengthening families programme 10–14: A randomized controlled trial. *European Journal of Public Health*, *26*, 953–959.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Barrera, M., Jr., & Castro, F. G. (2006). A heuristic framework for the cultural adaptation of interventions. *Clinical Psychology and Science Practice*, *13*, 311–316.

Bowen, M. (1991). Alcoholism as viewed through family systems theory and family psychotherapy. *Family Dynamics Addiction Quarterly*, *1*, 94–102.

Brody, G. H., Chen, Y.-F., Kogan, S. M., Yu, T., Molgaard, V. K., DiClemente, R. J., & Wingood, G. M. (2012). Family-centered program to prevent substance use, conduct problems, and depressive symptoms in Black adolescents. *Pediatrics*, *129*, 108–115.

Brody, G. H., Murry, V. M., Kogan, S. M., Gerrard, M., Gibbons, F. X., Molgaard, V., . . . Wills, T. A. (2006). The strong African American families program: A cluster-randomized prevention trial of long-term effects and a mediational model. *Journal of Consulting and Clinical Psychology*, *74*, 356–366.

Bröning, S., Baldus, C., Thomsen, M., Sack, P., Arnaud, N., & Thomasius, R. (2017). Children with elevated psychosocial risk load benefit most from a family-based preventive intervention: Exploratory differential analyses from the German strengthening families program 10–14 adaptation trial. *Prevention Science*, *18*, 932–942.

Brook, J., Akin, B. A., Lloyd, M., Bhattarai, J., & McDonald, T. P. (2016). The use of prospective versus retrospective pretests with child-welfare involved families. *Journal of Child and Family Studies*, *25*, 2740–2752.

Brook, J., McDonald, T. P., & Yan, Y. (2012). An analysis of the impact of the Strengthening Families Program (SFP 3–5 and 6–11) on family reunification in child welfare. *Children and Youth Services Review*, *34*, 691–695.

Burchinal, M., Follmer, A., & Bryant, D. (1996). The relations of maternal social support and family structure with maternal responsiveness and child outcomes among African-American families. *Developmental Psychology*, *32*, 1073–1083.

Castro, F. G., Barrera, M., & Martinez, C. R. (2004). The cultural adaptation of prevention interventions: Resolving tensions between fidelity and fit. *Prevention Science*, *5*, 41–45.

Chang, R., & Little, T. D. (2018). Innovations for evaluation research: Multiform protocols, visual analog scaling, and the retrospective pretest posttest design [Special issue]. *Evaluation and the Health Professions*. Design and method issues in drug use/abuse research: Special considerations and new approaches.

Cicchetti, D., & Rogosch, F. A. (1996). Equifinality and multifinality in developmental psychopathology. *Development and Psychopathology*, *8*, 597–600.

Coatsworth, J. D., Duncan, L. G., Berrena, E., Bamberger, K. T., Loeschinger, D., Greenberg, M. T., & Nix, R. L. (2014). The mindfulness-enhanced strengthening families program: Integrating brief mindfulness activities and parent training within an evidence-based prevention program. *New Directions for Youth Development*, *142*, 45–58.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*, 505–528.

DeMarsh, J. P., & Kumpfer, K. L. (1985). Family-oriented interventions for the prevention of chemical dependency in children and adolescence. *Journal of Children in Contemporary Society: Advances in Theory and Applied Research*, *18*, 117–151.

Dishion, T. J., Kavanagh, K., Schneiger, A., Nelson, S., & Kaufman, N. K. (2002). Preventing early adolescent substance use: A family-centered strategy for the public middle school. *Prevention Science*, *3*, 191–201.

Dishion, T. J., Nelson, S. E., & Kavanagh, K. (2003). The family check-up with high-risk young adolescents: Preventing early-onset substance use by parent monitoring [Special issue]. *Behavior Therapy*, *34*, 553–571.

Dishion, T. J., Shaw, D., Connell, A., Gardner, F., Weaver, C., & Wilson, M. (2008). The family check-up with high-risk indigent families: Preventing problem behavior by increasing parents' positive behavior support in early childhood. *Child Development*, *79*, 1395–1414.

Eiden, R. D., Lessard, J., Colder, C. R., Livingston, J., Casey, M., & Leonard, K. E. (2016). Developmental cascade model for adolescent substance use from infancy to late adolescence. *Developmental Psychology*, *52*, 1619–1633.

Elliott, D. S., & Mihalic, S. (2004). Issues in disseminating and replicating effective prevention programs. *Prevention Science*, *5*, 47–53.

Ferrer-Wreder, L., Adamson, L., Kumpfer, K. L., & Eichas, K. (2012). Advancing intervention science through effectiveness research: A global perspective. *Child and Youth Care Forum*, *41*, 109–117.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., … Li, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175.

Forehand, R. L., & McMahon, R. J. (1981). *Helping the noncompliant child: A clinician's guide to parent training*. New York, NY: Guilford Press.

Foxcroft, D. R., Callen, H., Davies, E. L., & Okulicz-Kozaryn, K. (2017). Effectiveness of the strengthening families programme 10–14 in Poland: Cluster randomized controlled trial. *European Journal of Public Health*, *27*, 494–500.

Foxcroft, D. R., Ireland, D., Lister-Sharp, D. J., Lowe, G., & Breen, R. (2003). Longer-term primary prevention for alcohol misuse in young people: A systematic review. *Addiction*, *98*, 397–411.

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*, 1261–1267.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translation Medicine*, *8*, 1–6.

Gorman, D. M. (2005). Drug and violence prevention: Rediscovering the critical rational dimension of evaluation research. *Journal of Experimental Criminology*, *1*, 39–62.

Gorman, D. M. (2017). The decline effect in evaluations of the impact of the Strengthening Families Program for Youth 10–14 (SFP 10–14) on adolescent substance use. *Children and Youth Services Review*, *81*, 29–39.

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

Guerney, B. G., Coufal, J., & Vogelson, E. (1981). Relationship enhancement versus a traditional approach to therapeutic/preventative/enrichment parent-adolescent programs. *Journal of Consulting and Clinical Psychology*, *49*, 927–939.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*, e1002106.

Heritier, S. R., Gebski, V. J., & Keech, A. C. (2003). Inclusion of patients in clinical trial analysis: The intention-to-treat principle. *MJA*, *179*, 438–440.

Hill, R. B. (1972). *The strengths of black families*. New York, NY: Emerson Hall.

Jessor, R., & Jessor, S. L. (1977). *Problem behavior and psychosocial development*. New York, NY: Academic Press.

Johnson-Motoyama, M., Brook, J., Yan, Y., & McDonald, T. (2013). Cost analysis of the strengthening families program in reducing time to family reunification among substance affected families. *Children and Youth Services Review*, *35*, 244–252.

Kameoke, V. A. (1996). *The effects of a family-focused intervention on reducing risk for substance abuse among Asian and Pacific-island youths and families:*

Evaluation of the strengthening Hawaii's families project. Honolulu: University of Hawaii, Social Welfare Evaluation and Research Unit.

Kogan, S. M., Lei, M.-K., Brody, G. H., Futris, T. G., Sperr, M., & Anderson, T. (2016). Implementing family-centered prevention in rural African American communities: A randomized effectiveness trial of the strong African American families program. *Prevention Science*, *17*, 248–258.

Kumpfer, K. L., Alvarado, R., Smith, P., & Bellamy, N. (2002). Cultural sensitivity and adaptation in family-based prevention interventions. *Prevention Science*, *3*, 241–246.

Kumpfer, K. L., Alvarado, R., Tait, C., & Turner, C. (2002). Effectiveness of school-based family and children's skills training for substance abuse prevention among 6–8 year old rural children. *Psychology of Addictive Behaviors*, *16*, S65–S71.

Kumpfer, K. L., & DeMarsh, J. P. (1985). Family environmental and genetic influences on children's future chemical dependency. *Journal of Children in Contemporary Society: Advances in Theory and Applied Research*, *18*, 49–91.

Kumpfer, K. L., Greene, J. A., Allen, K. C., & Miceli, F. (2010). Effectiveness outcomes of four age version of the strengthening families program in statewide field sties. *Group Dynamics: Theory, Research and Practice*, *14*, 211–229.

Kumpfer, K. L., Magalhães, C., Whiteside, H., & Xie, J. (2016). Strengthening families for middle/late childhood. In M. Van Ryzin, K. L. Kumpfer, G. Fosco, & M. Greenberg (Eds.), *Family-centered prevention programs for children and adolescents: Theory, research, and large-scale dissemination* (pp. 68–85). New York, NY: Psychology Press.

Kumpfer, K. L., Magalhães, C., & Xie, J. (2017). Cultural adaptation and implementation of family evidence-based interventions with diverse populations. *Prevention Science*, *18*, 649–659.

Kumpfer, K. L., Molgaard, V., & Spoth, R. (1996). The Strengthening Families Program for the prevention of delinquency and drug use. In R. D. Peters, & R. J. McMahon (Eds.), *Preventing childhood disorders, substance abuse, and delinquency* (pp. 241–267). Thousand Oaks, CA: Sage.

Kumpfer, K. L., Pinyuchon, M., de Melo, A., & Whiteside, H. (2008). Cultural adaptation process for international dissemination of the Strengthening Families Program (SFP). *Evaluation and the Health Professions*, *33*, 226–239.

Kumpfer, K. L., & Turner, C. (1990–1991). The social ecology model of adolescent substance abuse: Implications for prevention. *International Journal of the Addictions*, *25*, 435–463 (Original work published 1990).

Kumpfer, K. L., Whiteside, H. O., Ahearn Greene, J. A., & Cofrin-Allen, K. (2010). Effectiveness outcomes of four age versions of the Strengthening Families

Program in statewide field sites. *Group Dynamics: Theory, Research, and Practice*, *14*, 211–229.

Kumpfer, K. L., Xie, J., & Hu, Q. (2011). Engendering resilience in families facing chronic adversity through family strengthening programs. In K. Gow & M. Celinksi (Eds.), *Wayfinding through life's challenges: Coping and survival* (pp. 461–483). New York, NY: Nova Science.

Kumpfer, K. L., Xie, J., & O'Driscoll, R. (2012). Effectiveness of a culturally adapted strengthening families program 12–16 years for high-risk Irish families. *Child Youth Care Forum*, *41*, 173–195.

Little, R. J., & Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods*, *3*, 247–259.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

Maguin, E., Safyer, A., Nochajski, T. H., DeWitt, D., MacDonald, S., & Kumpfer, K. (2004). The impact of a family-based alcohol prevention program on parenting behavior. *Alcoholism: Clinical and Experimental Research*, *28*, 873.

Marek, L. I., Brock, D.-J. P., & Sullivan, R. (2006). Cultural adaptations to a family life skills program: Implementation in rural Appalachia. *The Journal of Primary Prevention*, *27*, 113–132.

Markstrom, C. A., Marshall, S. K., & Tryon, R. J. (2000). Resiliency, social support, and coping in rural low-income Appalachian adolescents from two racial groups. *Journal of Adolescence*, *23*, 693–703.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? *American Psychologist*, *70*, 487–498.

Molgaard, V. K., Spoth, R. L., & Redmond, C. (2000). Competency training: The strengthening families program: For parents and youth 10–14. Juvenile Justice Bulletin. Retrieved September 29, 2017, from https://www.ncjrs.gov/pdffiles1/ojjdp/182208.pdf

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Okulicz-Kozaryn, K., & Foxcroft, D. R. (2012). Effectiveness of the strengthening families programme 10–14 in Poland for the prevention of alcohol and drug misuse: Protocol for a randomized controlled trial. *BMC Public Health*, *12*, 319.

Olds, D. L., Eckenrode, J., Henderson, C. R., Kitzman, H., Powers, J., Cole, R., . . . Luckey, D. (1997). Long-term effects of home visitation on maternal life course and child abuse and neglect. Fifteen-year follow-up of a randomized trial. *Journal of the American Medical Association*, *278*, 637–643.

Olds, D. L., Robinson, J., Pettitt, L., Luckey, D. W., Holmberg, J., Ng, R. K., … Henderson, C. R. (2004). Effects of home visits by paraprofessionals and by nurses: Age 4 follow-up results of a randomized trial. *Pediatrics*, *114*, 1560–1568.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semi continuous longitudinal data. *Journal of the American Statistical Association*, *96*, 730–745.

Onrust, S., & Bool, M. (2006). *Evaluatie van de cursus gezin aan bod: Nederlandse versie van het Strengthening Families Program (SFP) [Evaluation of cursus gezin aan bod: The distribution. Dutch adaptation of the Strengthening Families Program (SFP)*] [In Dutch]. Utrecht, the Netherlands: Trimbos Institute.

Oretega, E., Giannotta, F., Latina, D., & Ciairano, S. (2012). Cultural adaptation of the strengthening families program 10–14 to Italian families. *Child Youth Care Forum*, *41*, 197–212.

Orte, C., Ballester Brage, L., Torelló, O., de Vicente, D., & Mascaró, A. (2017). Cultural adaptation of family evidence-based interventions. Results of the Spanish adaptation of SFP12–16. Presentation at annual EU SPR conference, September 22, 2017, Vienna, Austria.

Pantin, H., Coatsworth, J. D., Feaster, D. J., Newman, F. L., Briones, E., Prado, G., … Szapocnik, J. (2003). Familias Unidas: The efficacy of an intervention to promote parental investment in Hispanic immigrant families. *Prevention Science*, *4*, 189–201.

Parra-Cardona, J. R., López-Zéron, G., Domenech-Rodriguez, M., Escobar-Chew, A. R., Whitehead, M. R., Sullivan, C. M., & Bernal, G. (2016). A balancing act: Integrating evidence-based knowledge and cultural relevance in a program of prevention parenting research with Latino/immigrants. *Family Process*, *55*, 321–337.

Patterson, G. R. (1982). *Coercive family processes*. Eugene, OR: Castalia Press.

Patterson, G. R., Forgatch, M. S., & DeGarmo, D. S. (2010). Cascading effects following intervention. *Development and Psychopathology*, *22*, 949–970.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London, England: Routledge and Keagan Paul.

Puffer, E. S., Annan, J., Sim, A. L., Salhi, C., & Betancourt, T. S. (2017). The impact of a family skills training intervention among Burmese migrant families in Thailand: A randomized controlled trial. *PLoS One, 12*, *e0172611*.

Redmond, C., Spoth, R., Shin, C., & Lepper, H. S. (1999). Modeling long-term parent outcomes of two universal family-focused preventive interventions: One-year follow-up results. *Journal of Consulting and Clinical Psychology*, *67*, 975–984.

Reeves, M. P., & Rhine, J B. (1943). *The Journal of Parapsychology*, *76*.

Reichardt, C. S. (2011). Evaluating methods for estimating program effects. *American Journal of Evaluation*, *32*, 246–272.

Resnicow, K., Soler, R. E., Braithwaite, R. L., Ahluwalia, J. S., & Butler, J. (2000). Cultural sensitivity in substance use prevention. *Journal of Community Psychology*, *28*, 271–290.

Riesch, S. K., Brown, R. L., Anderson, L. S., Wang, K., Canty-Mitchell, J., & Johnson, D. L. (2012). Strengthening families program (10–14): Effects on the family environment. *Western Journal of Nursing Research*, *34*, 340–376.

Robling, M., Bekkers, M., Bell, K., Butler, C. C., Cannings-John, R., Channon, S., . . . Torgerson, D. (2016). Effectiveness of a nurse wellness-led intensive home-visitation programme for first-time teenage mothers (building blocks): A pragmatic randomised controlled trial. *The Lancet*, *387*, 146–155.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331.

Rutter, M. (2006). The promotion of resilience in the face of adversity. In A. Clarke-Stewart, & J. Dunn (Eds.), *The Jacobs Foundation series on adolescence. Families count: Effects on child and adolescent development* (pp. 26–52). New York, NY: Cambridge University Press.

Sanders, M. R., Baker, S., & Turner, K. M. T. (2012). A randomized controlled trial evaluating the efficacy of triple p online with parents of children with early-onset conduct problems. *Behaviour Research and Therapy*, *50*, 675–684.

Segrott, J., Holliday, J., Rothwell, H., Foxcroft, D., Muphy, S., Scourfild, J., . . . Moore, L. (2014). Cultural adaptation and intervention integrity: A response to Skärstrand, Sundell and Andreasson [In Swedish]. *European Journal of Public Health*, *24*, 354–355.

Segrott, J., Murphy, S., Rothwell, H., Scourfield, J., Foxcroft, D., Gillespie, D. , . . . Moore, L. (2017). An application of extended normalisation process theory in a randomized controlled trial of a complex social intervention: Process evaluation of the strengthening families programme (10–14) in Wales, UK. *SSM—Population Health*, *3*, 255–265.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Skärstrand, E., Larsson, J., & Andréasson, S. (2008). Cultural adaptation of the strengthening families programme to a Swedish setting. *Health Education*, *108*, 287–300.

Skärstrand, E., Sundell, K., & Andréasson, S. (2014). Evaluation of a Swedish version of the strengthening families programme. *European Journal of Public Health*, *24*, 578–584.

Small, S. A., Cooney, S. M., & O'Connor, C. (2009). Evidence-informed program improvement: Using principles of effectiveness to enhance the quality and impact of family-based prevention programs. *Family Relations*, *58*, 1–13.

Spoth, R., Randall, G. K., Shin, C., & Redmond, C. (2005). Randomized study of combined universal family and school preventive interventions: Patterns of long-term effects on initiation, regular use, and weekly drunkenness. *Psychology of Addictive Behaviors*, *19*, 372–381.

Spoth, R., Randall, G. K., Trudeau, L., Shin, C., & Redmond, C. (2008). Substance use outcomes 51/2 years past baseline for partnership-based, family-school preventive interventions. *Drug and Alcohol Dependence*, *96*, 57–68.

Spoth, R., & Redmond, C. (2002). Project family prevention trials based in community-university partnerships: Toward scaled-up preventive interventions. *Prevention Science*, *3*, 203–221.

Spoth, R., Redmond, C., & Shin, C. (2001). Randomized trial of brief family interventions for general populations: Adolescent substance use outcomes 4 years following baseline. *Journal of Consulting and Clinical Psychology*, *69*, 627–642.

Spoth, R., Redmond, C., Shin, C., & Azevedo, K. (2004). Brief family intervention effects on adolescent substance initiation: School-level growth curve analyses 6 years following baseline. *Journal of Consulting and Clinical Psychology*, *72*, 535–542.

Spoth, R., Redmond, C., Trudeau, L., & Shin, C. (2002). Longitudinal substance initiation outcomes for a universal preventive intervention combining family and school programs. *Psychology of Addictive Behaviors*, *16*, 129–134.

Spoth, R., Shin, C., Guyll, M., Redmond, C., & Azevedo, K. (2006). Universality of effects: An examination of the comparability of long-term family intervention effects on substance use across risk-related subgroups. *Prevention Science*, *7*, 209–224.

Spoth, R., Trudeau, L., Guyll, M., Shin, C., & Redmond, C. (2009). Universal intervention effects on substance use among young adults mediated by delayed adolescent substance use initiation. *Journal of Consulting and Clinical Psychology*, *77*, 620–632.

Spoth, R., Trudeau, L., Shin, C., & Redmond, C. (2008). Long-term effects of universal preventive interventions on prescription drug misuse. *Addiction*, *103*, 1160–1168.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71.

Tein, J.-Y., Mazza, G. L., Gunn, H. J., Kim, H., Stuart, E. A., Sandler, I., & Wochik, S. (2018). Propensity score approach to evaluating an effectiveness trial of the new beginnings program [Special issue]. *Evaluation and the Health Professions*.

Design and method issues in drug use/abuse research: Special considerations and new approaches.

Tryon, W. W. (2016). Replication is about effect size: Comment on Maxwell, Lau, and Howard (2015). *American Psychologist*, *71*, 236–237.

U.S. Department of Health and Human Services (HHS), Office of the Surgeon General. *Facing addiction in America: The Surgeon General's report on alcohol, drugs, and health*. Washington, DC: HHS; 2016. Retrieved from http://www.surgeongeneral.gov/library/reports/OR; https://addiction.surgeongeneral.gov

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., & . . . S. P. (2011). Replication in prevention science. *Prevention Science*, *12*, 103–117.

Van Ryzin, M., Kumpfer, K. L., Fosco, G., & Greenberg, M. (Eds.). (2016). *Family-based prevention programs for children and adolescents: Theory, research, and large-scale dissemination*. New York, NY: Psychology Press.

Wandersman, A., Duffy, J., Flaspohler, P., Noonan, R., Lubell, K., & Stillman, L. (2008). Bridging the gap between prevention research and practice: The interactive systems framework for dissemination and implementation. *American Journal of Community Psychology*, *37*, 384–399.