Article

Evaluation Review
2017, Vol. 41(5) 436-471

DOI: 10.1177/0193841X15625289
journals.sagepub.com/home/erx
SAGE

Interaction of Theory and Practice to Assess External Validity

Laura C. Leviton[1] and Mathew D. Trujillo[1]

Abstract

Background: Variations in local context bedevil the assessment of external validity: the ability to generalize about effects of treatments. For evaluation, the challenges of assessing external validity are intimately tied to the translation and spread of evidence-based interventions. This makes external validity a question for decision makers, who need to determine whether to endorse, fund, or adopt interventions that were found to be effective and how to ensure high quality once they spread. **Objective**: To present the rationale for using theory to assess external validity and the value of more systematic interaction of theory and practice. **Methods**: We review advances in external validity, program theory, practitioner expertise, and local adaptation. Examples are provided for program theory, its adaptation to diverse contexts, and generalizing to contexts that have not yet been studied. The often critical role of practitioner experience is illustrated in these examples. Work is described that the Robert Wood Johnson Foundation is supporting to study treatment variation and context more systematically. **Results**: Researchers and developers generally see a limited range

[1] The Robert Wood Johnson Foundation, Princeton, NJ, USA

Corresponding Author:
Laura C. Leviton, The Robert Wood Johnson Foundation, Rt. 1 and College Road East, Box 2316, Princeton, NJ 08543, USA.
Email: llevito@rwjf.org

of contexts in which the intervention is implemented. Individual practitioners see a different and often a wider range of contexts, albeit not a systematic sample. Organized and taken together, however, practitioner experiences can inform external validity by challenging the developers and researchers to consider a wider range of contexts. Researchers have developed a variety of ways to adapt interventions in light of such challenges. **Conclusions**: In systematic programs of inquiry, as opposed to individual studies, the problems of context can be better addressed. Evaluators have advocated an interaction of theory and practice for many years, but the process can be made more systematic and useful. Systematic interaction can set priorities for assessment of external validity by examining the prevalence and importance of context features and treatment variations. Practitioner interaction with researchers and developers can assist in sharpening program theory, reducing uncertainty about treatment variations that are consistent or inconsistent with the theory, inductively ruling out the ones that are harmful or irrelevant, and helping set priorities for more rigorous study of context and treatment variation.

Keywords
methodological development, outcome evaluation (other than economic evaluation), program implementation, program theory, real-world dissemination

Introduction

The external validity of evidence-based interventions (EBIs) is closely allied to translation from studies of effectiveness to field application and more widespread use of the EBIs. Indeed, there is strong consensus that weak external validity inferences greatly impede the spread of evidence-based and evidence-informed programs (e.g., International Initiative for Impact Evaluation, 2015). In this article, we will argue for, and illustrate, two ingredients that are essential to a rigorous and practical assessment of external validity: more adept use of program theory and more meaningful and systematic interaction between researcher–developers and practitioners.

To some, these may sound like platitudes: In principle we care about rigorous use of theory and including practitioners in evaluation. In reality, however, these present chronic challenges both for assessing external

validity and translating research to practice. Context and treatment variations bedevil the assessment of external validity as well as quality control for the spread of EBIs. Without sharper understanding of theory, we are severely hampered in determining whether an EBI can generalize to those contexts and whether treatment variations in an EBI are permissible. Evaluators have emphasized the interaction of theory and practice for many years, but their suggestions are not systematic to set priorities for assessing external validity. As we will describe, practitioners—working not individually but together in a structured arrangement—can help to sharpen theory and uncover critical features of context that are both frequent and important. Practitioners can work with researcher–developers to identify treatment variations that are consistent with theory, thus increasing the applicability of EBIs and reducing uncertainty about treatment variations that are helpful, harmful, or irrelevant.

Theory is necessary to the very definition of external validity and to the process of induction that underlies it. A brief refresher and definition of terms offer essential arguments for what follows. Incorporated into those definitions, we shall begin to describe how a more systematic interaction of theory and practice can be useful to better assess external validity. Beginning with these definitions, we shall demonstrate that assessing external validity requires a multimethod, systematic program of inquiry.

Definition of External Validity

Campbell and Stanley (1966) defined it as knowledge about "to what populations, settings, treatment variables, and measurement variables can [the effect] be generalized?" (p. 5). Recently, Shadish, Cook, and Campbell (2002) elaborated on this definition as assessing "whether the causal relationship holds over variation in persons, settings, treatments, and measurement variables" (p. 20). These definitions incorporate theory in the sense that populations, settings, and treatments represent samples from larger classes of such variables. We are interested in creating a sampling frame, precisely because stakeholders hold theories about why those particular classes are important.

For example, a central concern for home visiting programs is to generalize from the mothers and children sampled in the existing studies to the larger class of mothers and children. Generalization about home visitation's effectiveness is hampered when subclasses within the larger class have not received adequate study, such as mothers and children in tribal communities (Del Grosso, Kleinman, Esposito, Mraz Esposito,

Sama-Miller, & Paulsell, 2014). It is not only the population characteristics that matter, but that they are embedded, or nested, within a bundle of context features: providers with varying characteristics, organizations that vary, larger systems (e.g., tribal health care vs. private or state supported) and environments that may help or hinder the effort. The bundle of context features interacts with the EBI to produce variations in treatment, so exact replication is unlikely except, perhaps, in the very simplest of interventions.

For these reasons, assessing external validity cannot be exclusively a problem of sampling or statistical adjustment. Rather, multimethod studies and programs of study on EBIs are needed that can incorporate rigorous assessment of effectiveness and representativeness but also an ongoing probe of context, informed by theory. The statistical advances represented in this special issue are extraordinarily important, but their use is hampered by the deficits of most available studies. A random selection of individuals and sites might greatly assist in generalization, but very few studies of EBIs provide such sampling (Stuart & Rhodes, **In Press**; Tipton, Hallberg, Hedges, & Chan, **In Press**; Tipton & Peck, **In Press**). And the challenges are often formidable when trying to infer back to populations of participants or sites from study samples of convenience (Stuart & Rhodes, **In Press**; Tipton & Peck, **In Press**). Even if more studies did select sites randomly, the sample of sites is usually small, limiting the possible covariate adjustments. Moreover, purposive sampling or covariate adjustment at the site level presumes that researchers know enough about important and prevalent context variables to begin with. As Tipton, Hallberg, Hedges, and Chan conclude, assessment of external validity depends on knowing the covariates that are related to relevant variation (**In Press**).

External Validity as Informing Decision-Maker Choices

When their knowledge is limited, decision makers require ways to reduce uncertainty about their choices of action (March, 1994). With this challenge in mind, Cronbach and Shapiro (1982) reframed external validity as providing information under conditions of uncertainty and risk. For national funders, the issue is to determine where and when to endorse or fund interventions that were found to be effective. Their uncertainty involves the populations, contexts, and variations in treatment that have not been studied directly; the risk is that society's resources might be wasted. For local decision makers, the uncertainty is whether to adopt an intervention in their own, very particular combination of populations, providers, service organizations, and systems. The risk locally is that an EBI might not be

effective in the particular situation or might not be well implemented locally. And as Shadish et al. (2002) point out, knowledge about external validity is inevitably limited, since resources for studies are limited, our knowledge about important variation is limited; and even if these limits did not exist, it is not feasible for effectiveness studies to sample every variation of potential interest in populations, context, and treatment.

The Logic of Induction

External validity relies on induction, not deduction as in randomized experiments' tests of effectiveness. The truth of inductive reasoning is probable, not certain (Copi, Cohen, & Flage, 2007). Thus, one can have a theory about the classification of birds that says, "All birds have feathers and wings" (or at least, vestigial ones). Having wings is a necessary but not a sufficient condition for being a bird, since bats and other creatures also have wings. In the same way, the core components of an intervention (see below) are necessary conditions for effectiveness. Induction helps us determine whether an EBI will probably have benefit or no benefit, given a specific combination of population, setting, provider, and implementation characteristics. At a minimum, induction helps us rule out specific combinations as ineffective.

In many service delivery programs, the logic is that target populations (e.g., students, people at risk of HIV infection, new mothers, or incarcerated people) are recruited in some way and become program participants. Intervention (e.g., education, HIV prevention, home visitation, or prisoner reentry) is provided that aims to produce both intermediate outcomes for participants (e.g., engagement in school, negotiation to use a condom, parenting skills, or job skills) and ultimate outcomes (e.g., academic mastery, reduced AIDS cases, better child development, or reduced recidivism). Participants, intervention providers, and the service organizations that house the programs are embedded within larger systems (or lack of systems), and all of these levels affect the degree of program implementation.

This logic allows us to generalize by ruling out some contexts inductively, without new controlled studies, but through assessments of coverage and of the strength and integrity of implementation. For example, coverage of a population may be biased or not comprehensive. No participation, no program. We rule it out inductively for populations that don't participate; it is not a bird without feathers or wings. Even more starkly, when a program is not implemented, or implemented in too weak a form, or without the core components, it too is ruled out. In fact, we can easily generalize by ruling out: no program, no

effect. Epstein and Klerman (2013) have made the same point in the context of falsifiable logic models: Programs that do not have the necessary core components are a priori not effective and not worth evaluating.

Five Inductive Principles of External Validity

The more difficult challenge lies in *ruling in* the many combinations of populations, providers, organizations, treatment variations, and systems, where an intervention might be effective. Ruling in is always a probabilistic exercise, uncertainty reduction for the decision maker. To extend generalizations in the absence of formal sampling and tests of effectiveness, scientists use five principles (Shadish, Cook, & Campbell, 2002). (1) Assessing surface similarity between what is studied and the target of generalization. (2) Ruling out irrelevancies, context attributes that do not change a generalization. (3) Identifying context attributes that limit generalization, as in our examples of limited participation and implementation failure. (4) Interpolating to unsampled values within the sample range and extrapolating beyond the sample range. (5) Causal explanation in which scientists develop and test explanatory theories about the target of generalization.

We will illustrate below how a better interaction of theory and practice can inform these principles. In doing so, we challenge the assumption that only "scientists use" them. As we will demonstrate, in real-world settings, it is often more useful if practitioners participate in using these principles. It is not only that practitioners have essential information to offer—it would be healthier if evaluators used more of their input, but that is not all. It is also their sense-making capacity that is needed because the contexts are at least complicated and often complex. In particular, practitioners have something to offer in dealing with treatment variation, and there are structured ways to do so.

Coping With Complexity: The Importance and Prevalence of Context Features

Writers on complexity in evaluation tend to assert that each context is unique (Patton, 2010; Pawson & Tilley, 1997). Yet they acknowledge that across contexts, we detect patterns in the chaos, and they are highly useful. The patterns are seen when we un-bundle context: providers, organizations, systems and environments that may produce variation in outcomes. In a program of inquiry about external validity (as opposed to a single study),

context features can be prioritized by two criteria that allow us to determine whether they are consequential for population impact:

- *Importance:* We can identify many regularities, classes of variables that, on the basis of observation and study, appear important to affect program outcomes. These can receive priority for further study to extend generalizations. Program theory guides decisions about importance (see below).
- *Prevalence or frequency:* Writers on context are correct that the various combinations of these variables are daunting, given the number of statistical interactions that are possible. Yet certain combinations of population, setting, provider, and so forth, are quite simply more prevalent or frequent than others.

For example, early intervention services for infants and toddlers with developmental delays were primarily delivered at home in 2006, with only 5% delivered at clinics or early intervention centers (Johnson, 2009). Caucasian, Latinos, and African American families are simply more plentiful than other ethnicities in the United States. While no one wants to ignore other populations in need, or service settings, the high frequency combinations have to take initial priority for assessing external validity (unless there is another compelling reason). As we will illustrate, assessments of prevalence, or at least frequency of such combinations, offer a guide to assess external validity more efficiently.

"Reach" is a helpful concept for this purpose: It is defined as the proportion of intended beneficiaries that can be exposed to an intervention (Glasgow, Vogt, & Bolles, 1999). While similar in some ways to Rossi, Lipsey, and Freeman's (2004) concept of coverage, reach is intentional about program design: recruitment, engagement, and ensuring that the intervention is appropriate for the target group. The original external validity definition does not, strictly speaking, include reach, but both policy makers and practitioners want to ensure reach.

Summary

External validity acknowledges the fact of variation, which increases uncertainty for both decision makers and practitioners about where an intervention will be effective, for whom, and in what context. Because the number of combinations of populations, settings, practitioners, and treatment variations is endless, external validity is always an inductive process. Program

theory is essential to identify the regularities across contexts where an EBI is used. Then, the populations, settings, and other context features that have a greater importance and frequency can take priority for assessment. Throughout the rest of this article, we will illustrate how researchers use theory to assess external validity, and how a more meaningful exchange with practitioners extends the power of the inductive process.

Theory as a Foundation of External Validity

A Brief Refresher on Theory

People naturally hold theories, whether they are implicit or explicit and formal or informal. Theory and practice represent different levels of abstraction; their interaction is essential to applied fields like evaluation (Leviton, 2015). Writers on evaluation increasingly value practical knowledge in addition to measured knowledge and assert that practice and theory are intertwined (Stake & Schwandt, 2006). Ostrom's (1990, pp. 45–46) description is fundamental to the strategies we will describe for assessing external validity:

> Understanding how individuals solve particular problems in field settings requires a strategy of moving back and forth from the world of theory to the world of action. Without theory, one can never understand the general underlying mechanisms that operate in many guises in different situations. If not harnessed to empirical problems, theoretical work can spin off under its own momentum, reflecting little of the empirical world.

Theory helps to sharpen what fidelity and the essential program components are really all about; thus it helps with quality control in implementing EBIs. More generally, theory helps us to assess what is knowable and predictable in complex, or at least complicated, systems of service delivery. Without theory, or at least some abstraction of concepts, we are left with an unrealistically narrow and inflexible range of program activities and contexts to assess external validity.

Although philosophers of science hold sometimes conflicting views about theory, Hasok Chang (2012) concludes that theory's primary function is usefulness. We subscribe to this view, since the primary justification for evaluation is also its presumed usefulness (Shadish, Cook, & Leviton, 1991). Chang reviews the history of science to support his view that pluralism about theory is more realistic and useful than is monotheory science. This argument is important for external validity, given the many factors that

operate at different levels on a given social problem. Combining behavioral theories has proven useful for EBIs (e.g., Michie, West & Campbell, 2014), and combining organization and systems theories helps to assess capacity to implement an EBI (Schuh & Leviton, 2006). It is permissible to combine theories where the combination is useful and theories are not otherwise in conflict. Criteria exist for whether program theories are good or bad, consistent or contradictory, useful or not useful, and well tested or unsupported (Davidoff, Dixon-Woods, Leviton, & Michie, 2015).

Three general types of theory are useful to inform external validity: descriptive, causal, and explanatory (Davidoff et al., 2015). Descriptive theory applies, for example, when cultural groups are described as sharing a common history and expectations. In the context of programs, one quickly moves to causal theory, which asserts a causal relationship between two or more variables, as between an intervention and its impact on a social problem. Causal theory is amenable to test: for example, whether in some programs, the tailoring of interventions to cultural characteristics (often a deliberate adaptation of an original program model) results in bigger effect sizes or improved reach to benefit entire populations (Castro, Barrera, & Holleran Steiker, 2010). In some programs, cultural tailoring improves effect sizes (e.g., Griner & Smith, 2006) and in others, it does not (e.g., Robinson et al., 2010). Explanatory theories address such mixed findings by exploring and analyzing the mediating variables that provide a mechanism for the tailored intervention to achieve its results, or the population and provider characteristics known to moderate the effects of the original, unadapted intervention (e.g., Jagers, Syndor, Mouttapa, & Flay, 2007).

Theory-Based Evaluation Approaches

Program evaluation borrows from social science's grand theories (applicable over many content areas), and the middle range theories (delimited in their areas of application), but primarily focuses on the specific "small theory" about how a particular program is supposed to work (Lipsey, 1993). Writers such as Chen and Rossi (1992), Flay (e.g., Flay, Berkowitz, Bier, & The Social and Character Development Research Consortium, 2009), and Weiss (1997) have long advocated the use of program theory to plan and interpret evaluation studies in areas as diverse as criminal justice, health, education, and social welfare. Theory helped these evaluators to better specify models of intervention and to test the effects of mediating and moderating variables. On occasion, detail about the underlying program theory assisted users to detect important patterns in systematic

reviews (Lipsey 2009) or plan strong tests of a program theory that could rule out alternative explanations (e.g., Lipsey, 1993). Implementation science (Eccles & Mittman, 2006; Fixsen, 2015) extends the use of theory beyond characteristics of the program recipients and interventions to the sometimes critical importance of context features: service provider characteristics, organizational capacity, and features of the broader systems in which they are nested.

Activities, Program Components, and Underlying Theory

Model developers and researchers speak about core components, essential elements, and key principles or functions of the program activities. Although hairs are split, these synonyms all refer to an underlying theory of how the intervention is supposed to be delivered (Lipsey, 1993; Weiss, 1997). Program activities make the underlying principles operational. In Cronbach and Meehl's (1955) terms, the observable activities are representations of more abstract core components, which together represent theoretical constructs. To assure that core components are provided, practitioner training and program manuals of operation specify the sequence and time spent on activities. Departures from specifications may water down effects by adversely affecting both the integrity and strength of interventions (their intensity, duration, and relevance, Yeaton & Sechrest, 1981). It should be noted that for some developers, the specified activities and core components are identical, making any adaptation highly problematic.

Generalization About Theoretical Constructs Identified Through Systematic Reviews

Theoretical constructs that are essential to effectiveness can sometimes be identified through bodies of evidence. A case in point is Lipsey's (2009) meta-analysis that determined a positive effect of therapeutic interventions for juvenile delinquency and a negative effect of punitive ones. The interventions shared, and reported, certain features that allowed them to be distinguished in these two categories. What the interventions did not share was a manual of operations or training process. Rather, the two categories derived from competing theories of how to address juvenile delinquency, and theory also explains why punitive interventions would be harmful (e.g., Center for Mental Health in Schools at University of California, Los Angeles, 2008).

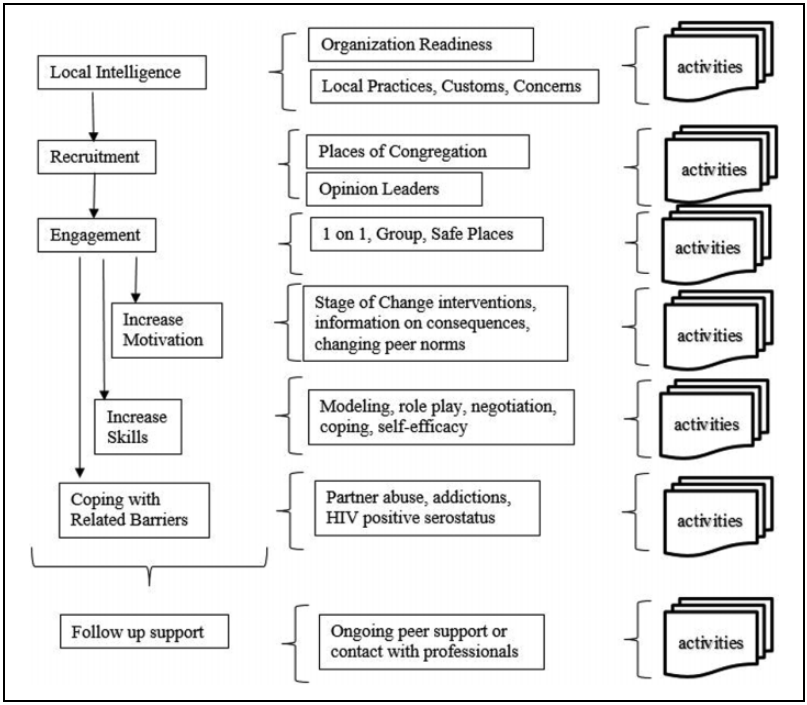

Figure 1. Greatly oversimplified flow chart of program components for HIV prevention.

An Illustration of Core Components and Underlying Theory

Primary prevention of HIV infection gives some of the most elaborated examples of theoretical constructs and core components made operational—adapted—in many different ways. The effectiveness of these program components has been tested with diverse groups at risk across many settings but usually in the context of multifaceted programs (Centers for Disease Control and Prevention [CDC], 2014). Figure 1 illustrates a family of interventions that have been found to be effective. It is a flowchart of components shared across many models, inferred from the extensive literature (Leviton & Guinan, 2003); it is necessarily a sketch and incomplete. Theoretical constructs are listed to the left, core components are in the middle, and a range of specific activities that could make them operational appear on the right. Intervention often includes:

(1) *Local intelligence:* Assessment of the local situation to assure organizational readiness to implement and an understanding of the local context, in order to know where and how recruit people at risk and to make interventions relevant to their concerns. (2) *Recruitment:* Contact with presumed peers or opinion leaders in progam storefronts or at people's places of congregation. (3) *Engagement* of people at risk, most often including (a) tailoring to the stage of change for reducing risky behaviors, motivational interviewing, and persuasive efforts to personalize the risk and support peer norms for prevention; and (b) skills training for those who are ready to protect themselves, for which mastery may increase a sense of control. (4) *Coping with related barriers:* For especially vulnerable groups, tailored counseling on issues that may affect their ability to protect themselves, such as partner abuse, addictions, or HIV-positive serostatus. (5) *Follow-up support:* for some programs, ongoing availability of supportive peers, booster sessions, or a relationship with a professional.

This sketch illustrates the principle that diverse activities may well be "many roads to Rome" to achieve the purposes of core components and the underlying constructs. For example, gay and bisexual men originally participated through formal counseling sessions but reach was limited, so formats were created for gay bars. Other populations at risk were then tackled, requiring street outreach for injection drug users (IDUs), "safe places" in public housing for women at risk, and many other variations (Leviton & Guinan, 2003). As long as the venues were helpful in recruiting people at risk, the specific recruitment activities could vary. The sketch also gives hints about activities that are not compatible with these programs. Thus, people at risk could be forcibly recruited through health departments' legal powers; however, such a practice may backfire given the importance of trust and peer engagement to the later program components.

The components of engagement have likewise been operationalized in many ways. Among other theories, the engagement process relies on the Theory of Reasoned Action, which specifies three factors associated with behavior change: motivation to avoid the risk, changing people's perceptions of peer norms, and giving them a sense of control. While the relative importance of these three factors varies from one situation to another, empirically they influence behavior (Fishbein & Ajzen, 2010). Changes in these three factors constitute intermediate outcomes. To achieve them, the interventions make intentional use of relevant local concerns, adapting materials and skills training to local circumstances.

Summary

Theory is at the heart of EBIs, in that it specifies the core components that make EBIs effective. However, as seen in the HIV example, these core components can be made operational in a variety of ways, so that a family of activities all fall under the treatment constructs and core components. The family of activities is not the same as a manual of operations that specifies exactly what must be done in an EBI. Treatment constructs make it abundantly clear that there is room within EBIs for reasonable adaptation of core components. HIV prevention also owes a continuing debt of gratitude to practitioners, who sharpened the core components underlying these approaches through their interactions with researchers.

Why an Interaction With Practitioners Is Essential to Assess External Validity of EBIs

There are three arguments for greater participation of practitioners in assessing external validity. First, as Ostrom (1990) pointed out, theory is sterile without action and reality testing. Second, variation in treatments needs more serious assessment. Local adaptation—treatment variation—appears inevitable, so it requires better study. Third, practitioners and researchers both have knowledge to contribute to sharpen theory and improve the process of induction.

EBIs Need More Reality Testing

The Canadian Institutes of Health (2014) define research translation as "a dynamic and iterative process that includes the synthesis, dissemination, exchange and ethically sound application of knowledge to improve ... services and products ..." (p. 1). However, in the United States, there are severe limits on any "exchange" of knowledge, in terms of understanding the variety of contexts where interventions are delivered. From our discussions with federal staff in education, prevention, and social welfare, national initiatives rarely solicit proactive, systematic feedback from implementers. It is often up to the original developers or researchers to identify populations, providers, organizations, and systems, where they want to assess external validity, overcome barriers, take advantage of facilitators, or otherwise modify interventions in light of new situations. A scan by the CDC indicated that developers and researchers vary greatly in terms of their

time, inclination, and skill to deal with new contexts where their interventions might be applied (Perkinson, 2012).

Even with the best will in the world, developers and researchers cannot anticipate all the contexts where adaptation might be necessary. Simply by the numbers, practitioners *taken together* are likely to see a wider range of contexts than do the developers or researchers, and we have already pointed to the prevailing limitations of study samples for assessing external validity. Even programs with widespread application can encounter new contexts. For example, the transitional care model (TCM) provides comprehensive in-hospital planning and home follow-up for chronically ill high-risk older adults. TCM is one of the best tested interventions in health services research, has undergone extensive testing to identify the core program components, and through widespread adoption has identified important facilitators and barriers to implementation (Naylor et al., 2009). Nevertheless, TCM's study of local adaptation began with a systematic search to identify unknown users. The investigators received hundreds of responses from many sources, along with many queries about how the model might be adapted in new circumstances (Mary Naylor, personal communication, November 09, 2015).

Treatment Variation Needs More Serious Assessment

Local adaptation is a well-known tendency in the diffusion of innovations (Rogers, 2003). Adaptations of EBIs depart from the originally specified activities and may even depart from an underlying logic model or program theory. Local adaptation is seen in fields as diverse as education and criminal justice (Emshoff et al., 1987), health promotion and disease prevention (Dusenbury, Brannigan, Hansen, Walsh, & Falco, 2005), mental health and social services (Chorpita, Becker, & Daleiden, 2007), quality improvement in medicine (Berwick, 2003), substance abuse (Peters & Wexler, 2005; Ringwalt et al., 2003), and violence prevention (Freire, Perkinson, Morrel-Samuels, & Zimmerman, 2015).

It is unclear how often adaptations occur and what their general consequences are. The concern is that adaptation may be an implementation failure and a departure from fidelity to the EBI. This is an appropriate concern, since departures from fidelity sometimes reduce program outcome effect sizes (e.g., Elliott & Mihalic, 2004; Hulleman & Cordray, 2009). Rossi's (1987) "Iron Law"—which he himself agreed was overstated—asserts that the expected effect of large-scale programs is zero, in part because of implementation failures. Programs implemented in open

systems do tend to have smaller effects than those implemented under more controlled research conditions (Wilson & Lipsey, 2001), and adaptation that is inconsistent with EBI specifications may be one cause. Yet a growing body of evidence indicates that adaptation can occur side-by-side with fidelity, as seen below (Backer, 2001; Dusenbury et al., 2005). Some adaptations are aligned with program goals and others are not, but this issue requires more study (Blakely et al., 1987). In fact, the available studies are mixed as to whether specific adaptations actually improve or harm outcomes (Bishop et al., 2014; Durlak & Dupre, 2008). Whether adaptations are helpful or harmful is an empirical issue—it is an evaluation issue (Rohrbach, Grana, Sussman, & Valente, 2006).

In many cases, local adaptation represents something different from the implementation failures caused by dishonesty and carelessness, of which evaluators have long been aware (Pressman & Wildavsky, 1984). In fact, local adaptation of EBIs occurs even when practitioners operate in good faith, with high capacity, and with a good understanding of program theory and specifications. As we will describe, local adaptation sometimes provides an opportunity to increase the quality and relevance of interventions, when implemented properly. The challenge comes in understanding what "proper implementation" may mean.

Practitioners and Researchers Have Complementary Knowledge for External Validity

As seen in Table 1, researchers and practitioners have different strengths and limitations for assessing external validity. The rows distinguish practitioners and researcher–developers, but critically important, they also distinguish novices from experts or reflective practitioners and researchers (Expertise is a continuum, suppressed in Table 1 for clarity and contrast). The columns distinguish knowledge of the EBI, further divided into knowledge of internal validity and of theory, contrasted with knowledge of practice, further divided into practice repertoire (the intervention skills one can bring into play), and the sample of practice contexts one knows about.

In any human activity, expertise varies (Dreyfus & Dreyfus, 1988). Novices require rules-based guidance, while experts process the world differently, just as a chess master processes the board differently than a beginner. Experience is not the same as expertise, however. Both experts and merely competent practitioners can have experience. Schön (1983) crystalized the difference in his term, reflective practice. Reflective practitioners draw on a wider range of experiences and have developed a larger repertoire

Table 1. What Scientists and Practitioners Know About External Validity.

	Knowledge of the EBI		Knowledge of Practice	
Role/Expertise	Internal Validity	Program Theory	Practice Repertoire	Sample of Practice Contexts
Novice practitioner	Limited ability to test effectiveness outside experimental conditions	Limited: fidelity concerns greatest	Limited range of experiences; total reliance on rule book or manual	Limited: fidelity concerns the greatest
Expert, reflective practitioner	Limited ability to test effectiveness outside experimental conditions	Potentially extensive: depends on length of experience with the EBI	Extensive: insights on adaptation	Extensive: • Strength: real-world challenges and opportunities • Weakness: sample is selective
Novice researcher–developer	Variable	Limited: • underspecified theory • detail lacking on core components	Limited: • few interactions with practitioners • doctrinaire about adaptations	Limited: • Strength: controlled studies • Limitation: few contexts
Expert, reflective researcher–developer	Often extensive, including explanatory analyses (mediators and moderators)	Extensive: • theory highly articulated • many ways to operationalize core components	Potentially extensive: insights on adaptation if experience with practitioners	Potentially extensive: • Strengths: controlled studies and systematic sample of contexts • Weaknesses: ability to sample adaptations and challenges and opportunities

Note. EBI = evidence-based intervention. Expertise is a continuum, suppressed for clarity and contrast.

of responses than have novices or merely competent practitioners. In new situations, they draw on experience, then assess the results and incorporate them into their repertoire. Expert practitioners have a broad and deep knowledge base; examples include master teachers, veteran therapists and social workers, and seasoned health-care providers or law enforcement personnel. For assessing external validity, expert practitioners have different things to say than do novices or merely competent practitioners.

Practitioner knowledge is based on both training and lived experience (Davidoff et al., 2015; Stake & Schwandt, 2006). Researchers often frame practitioner experience in terms of deficits: bias in their assessment of effectiveness (internal validity), the limited or biased sample of contexts individual practitioners have seen (representative sampling), or sometimes, their ability to make theory operational. However, practitioners' repertoires can be extensive and their recognition of real-world challenges and opportunities can be formidable. It depends on expertise. Novices will know little about practices of any kind, including EBIs. They need the manual of operations as well as extensive training and supervision. It is safe to assume that they understand the least about fidelity and will make the most mistakes—in itself this provides important information for external validity, especially for quality control and fidelity (e.g., Morrel-Samuels, Hutchison, Perkinson, Bostic, & Zimmerman, 2014; Rotheram-Borus, Swendeman, & Becker, 2014). In contrast, reflective practitioners can better contribute to assessing external validity and translating interventions, given the range of contexts that they have seen. Once theory is understood, their proficiency in making activities operational can also be remarkable: We have witnessed many examples in the HIV, medical care quality improvement, and health promotion areas. Reflective practitioners can help make sense of context, elaborate theory, adapt in sensible ways to overcome barriers, or work to increase the relevance and quality of intervention. We will present examples, subsequently.

Researcher–developers possess deficits as well as strengths, just as practitioners do. It may seem odd to say that researcher–developers can be novices or reflective experts, but that is clearly the case. Interventions that are termed EBIs vary in terms of the quality and volume of evidence, as seen in the varying standards of evidence used by national clearinghouses, the varying number of studies seen in systematic reviews, the lack of detail on context in those studies and systematic reviews (Avellar, Kleinman, Miller, et al., **In Press**; Horne, **In Press**), and the ongoing need to better specify what interventions consist of (e.g., Hoffmann et al., 2014). Researcher–developers have varying interest in addressing these problems.

In addition, many EBIs have seriously underspecified theory and lack detail on the core components, at least in the areas of violence prevention (Perkinson, 2012), mental health, and health promotion (Rotheram-Borus et al., 2014). Novice researcher–developers may or may not attend to the need to sharpen theory or may equate the EBI with the specific activities in a manual of operations. They are then indignant when practitioners say they cannot implement "by the book." Reflective researcher–developers on the other hand find many ways to operationalize the core components, have highly articulated theories underlying their programs, have studied a greater variety of contexts, and most important, they seem to listen to practitioners (see below).

Practitioners can challenge developers to reconsider and refine the underlying model and permissible activities, even develop a sharper program theory through these challenges. *Active Living Every Day*® is a classroom-based program based on principles of the transtheoretical model (Prochaska & Norcross, 2009) that encourages people over 50 to become more physically active. The Robert Wood Johnson Foundation (RWJF) supported an effort to implement the model in diverse populations (ethnicity, income, middle aged, vs. older) and organizations (e.g., senior centers, faith based, Young Men's Christian Association (YMCA), and healthcare plans). Some participants had Spanish as their first language, and practitioners mentioned that materials in Spanish would likely be more effective for them. Simple back translation was not sufficient; the cultural meaning of the materials also had to be incorporated. Practitioners also asked whether the materials could be concentrated into a smaller number of classroom sessions, which would attract more working people and also help sustain the program at their agencies. Concentrating the number of sessions made no difference to the outcome (Lattimore et al., 2010; Wilcox et al., 2008). It would be wrong to say the theory behind *Active Living Every Day* was underspecified. At the same time, the interaction with practitioners was a gift to the program by sharpening the core components and permissible variations.

Figure 2 may clarify some of the discussions of fidelity and adaptation and possibly inform external validity more generally. It represents challenges and opportunities given four combinations of novice and expert practitioners and researcher–developers, with deeper shading indicating greater challenges and lighter quadrants indicating the combinations that are most likely to contribute to generalizability and spread of EBIs. The dynamics in Figure 2 are merely suggestions, but they present likely hypotheses. In some quadrants, there are documented, excellent examples

Role	Novice Practitioners	Expert, Reflective Practitioners
Novice Researcher or New EBI	Most challenges in this quadrant: greatest needs are to designate core components and monitor fidelity	Policing of fidelity may be misplaced; de-skilling of the practitioner; loss of opportunities for application and relevance
Expert, Reflective Researcher or Long-established EBI	Clarify the manual of operations, training and TA, monitor fidelity but use feedback	Most opportunities in this quadrant: Systematic expansion of ways to make theory operational; ongoing development of EBI, permissible adaptations, and contexts for generalizability

Notes: Deeper shading indicates more problems. Continuum of expertise is suppressed for clarity and contrast.

Figure 2. External validity heat map.

(e.g., Rotheram-Borus et al., 2014, for the bottom two quadrants; quality improvement in medicine for the bottom right quadrant).

How Practitioners Contribute to the Five Inductive Principles of External Validity

The example of HIV prevention describes a family of adapted interventions that were already applied and empirically tested in various contexts. We believe on this basis that we can generalize to a larger class of interventions, consistent with theory. But what is the decision maker to do when practitioners want to implement an EBI with populations that have not been studied yet, with types of providers, organizations, and systems that differ, sometimes markedly, from those sampled in the available studies? In such situations, uncertainty is at its greatest. Shadish et al.'s (2002) five principles of generalization help to reduce this uncertainty. Practitioners can assist.

Surface Similarity and Irrelevancy: Who Recognizes Them as Such?

The *Active Living Every Day* example illustrates how practitioners contribute to assessing external validity, using Shadish et al.'s (2002) principles of surface similarity and irrelevancy. Properly translated, Spanish materials are similar to English materials, so one can probably generalize the program to Spanish speakers. And yet, the cultural meaning of the materials goes far beyond the scientist's expertise, so it is not really surface similarity at all! The similarity was only "surfaced" through practitioner interaction and careful adaptation. In the same way, the practitioners and developers together identified an irrelevancy: the originally specified number of classroom sessions. Paring back the number of sessions sometimes reduces the strength or integrity of intervention, but not in this case.

Context Attributes That Limit Generalization: Who Will Recognize Them?

Practitioners can also inform Shadish et al.'s (2002) third principle, around context attributes that limit generalization. In many programs, access is a problem, as when language translators are needed for patient instructions or when transportation is needed to get patients to the health providers in the first place. A dominant model of health-care access with decades of study supports the importance of these context attributes (Babitsch, Gohl, & von Lengerke, 2012), yet practitioners are likely to notice them in the first place, so that the

theory could be applied! Moreover, practitioners and developers can work together to overcome access problems and take advantage of opportunities that context may present. For example, HIV prevention with IDUs required adaptation of generic health education and skills training that had been found effective previously. Prevention for IDUs is more effective when they are recruited and educated by outreach workers who are themselves former IDUs. Such outreach workers are credible message sources, as specified by general communications theory (Hovland, Janis, & Kelly, 1953) and adapted to the context of a marginalized group. The effectiveness of this adaptation is amply supported (Coyle, Needle, & Normand, 1998).

A Detailed Example to Illustrate the Fourth and Fifth Principles

Several, diverse interventions aim to inculcate a sense of belonging in college students from marginalized backgrounds, so as to retain the students in academic settings and promote academic achievement. These interventions are predicated on the theory that belonging is a fundamental human motivation (Baumeister & Leary, 1995). While people in general want to feel that they belong, this feeling varies as a function of group membership. Members of stigmatized groups may feel uncertain about their belonging in mainstream institutions, including college. This feeling is a result of their marginalized status and associated concerns over being accepted by individuals from nonstigmatized groups. Individuals who feel uncertain about the extent to which they belong may become hypervigilant to environmental cues that signal whether they do not belong and it may lead them to interpret ambiguous events as evidence of their not belonging. As a result, feelings of uncertainty about belonging can quickly lead to feelings that one does not belong. The feeling that one does not belong, in turn, can cause high levels of stress and threat and lead to negative outcomes for individuals from marginalized groups (Steele, Spencer, & Aronson, 2002).

Feelings among individuals from marginalized groups that they do not belong to their institution of higher learning provide one causal explanation for disparities in educational outcomes (Steele et al., 2002). Students may feel that they will not be accepted by their teachers or peers from nonstigmatized groups. Students from stigmatized groups would then be constantly searching for evidence that they do not belong, making them interpret commonplace college experiences, such as having difficulty making friends or receiving critical feedback, as signs that they do not belong. This feeling that one does not belong would then negatively affect student's persistence in studies and therefore their academic performance. Accordingly, should

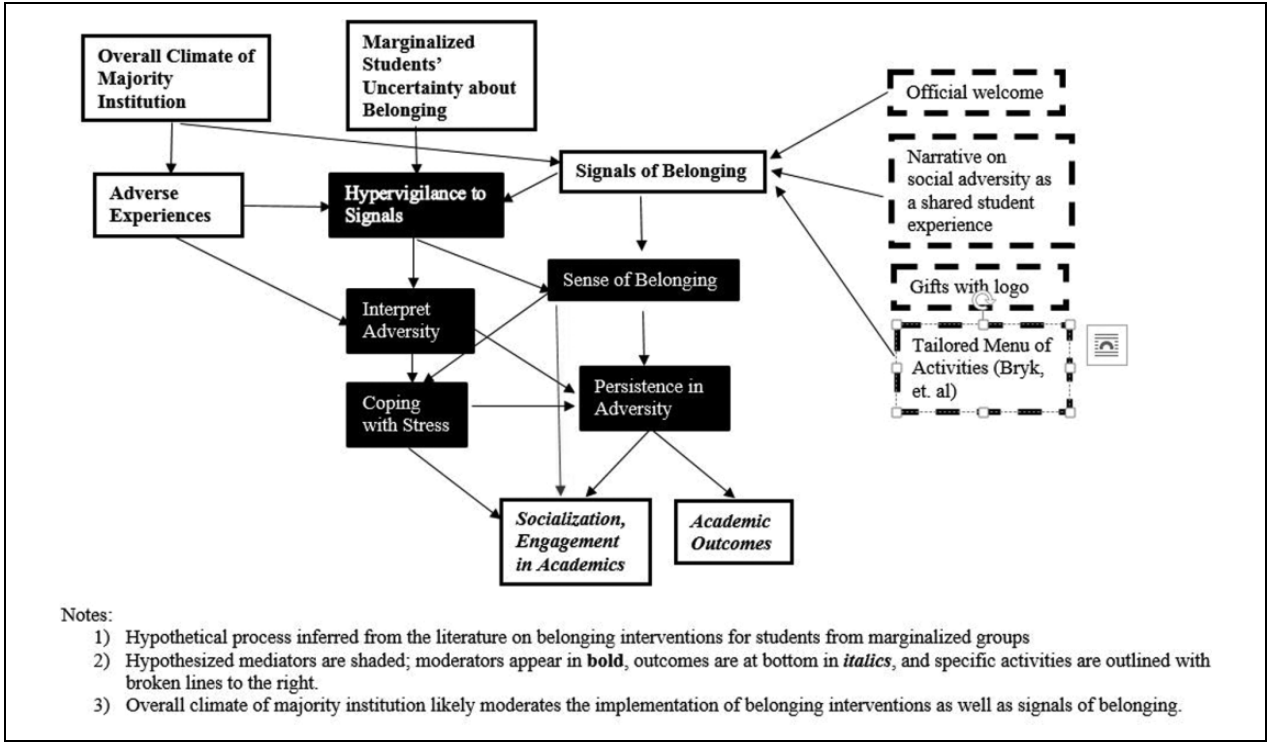

Figure 3. Mediators and moderators of belonging interventions in education.

one be able to intervene to decrease individual's lack of certainty that they belong, they may be better able to cope with the common adversity of education and may have improved educational outcomes.

Several "belonging" interventions are summarized in Figure 3, along with mediators and moderators of their effect on students. At a large public university, African American students' sense of belonging was a significant predictor of both their identification with the institution and intentions to persist (Hausmann, Schofield, & Woods, 2007). To improve these students' sense of belonging, they received multiple written communications from university administrators, indicating that they were valued members of the university community. The students also received small gifts that displayed the university's name, logo, and colors. Compared to those in the control condition, the students experienced less rapid decline in sense of belonging and intentions to persist over time. A second intervention took place at an elite private institution: African American students read a narrative that framed social adversity as an experience that is both short lived and shared across all students. Compared to control students, they reported a greater sense of belonging, potential, engagement in achievement behavior, and grade point average (GPA; Walton & Cohen, 2007). Moreover, the intervention mitigated the degree to which African American's sense of belonging varied as a function of their daily adversity. The same intervention at a second elite private institution caused a positive impact on GPA 3 years postintervention (Walton & Cohen, 2011).

Researchers have also applied this intervention to women in science, technology, engineering, and mathematics (STEM) fields. It is important to note that the intervention was adapted to incorporate aspects of belonging of specific relevance to women in STEM. For women in male-dominated STEM fields, this intervention led to higher GPA, daily functioning, felt experience with engineering, confidence in prospects in engineering, more positive implicit norms, and greater reporting of having male engineer friends (Walton, Logel, Peach, Spencer, & Zanna, 2015). Interestingly, the intervention was only effective for women in male dominated STEM fields—where they might not otherwise feel a sense of belonging—and not for women in STEM fields where the genders were represented more equally. In addition to testing this intervention in higher education, researchers have tested it among middle school students. The intervention alleviated African American student's uncertainty about social belonging and improved their GPA (Walton, Cohen, Garcia, Apfel, & Master, 2012).

Using Theory for Interpolation and Extrapolation

Our review of interventions to create a sense of belonging illustrates how one might interpolate and extrapolate to populations and settings that are not yet studied. Belonging interventions improved educational outcomes for students at both elite, private, 4-year universities and at community colleges, which are conceptually at opposite ends of a continuum of higher education institutions. Yet, belonging interventions also improved retention at a public 4-year university, conceptually between the two ends of the continuum. By interpolation to unsampled values that fall within a sampled range, one may generalize that the effect of these interventions may apply to the entire range of higher education institutions, provided that context attributes do not limit the generalization—for example, an overall climate of hostility to the marginalized group. Research on the relationship between a sense of belonging and educational outcomes also has utility for understanding the principle of extrapolation—generalizing about unsampled values that fall outside of a sampled range. Although much of the research on belonging interventions focuses on African American students, one can use the theory to generalize to other, unsampled, marginalized groups that express uncertainty about whether they belong in an academic setting. Once again, research supports this extrapolation—by indicating that belonging interventions also improve educational outcomes for women in male dominant STEM fields, but not in STEM fields where the genders were represented more equally. Practitioners can contribute to interpolation and extrapolation by suggesting situations and groups where belonging interventions are likely to be needed. The prevalence and importance of these situations and groups could provide a basis for additional study.

Using Explanatory Theory to Generalize About Belonging Interventions

The specific activities of these interventions were adapted to context. What remained constant were the underlying principles. Moreover, the mediators (e.g., increased feelings of belonging and positive coping with adversity) and outcomes (e.g., GPA) were fairly consistent across these studies, leading to the conclusion that they fall into a larger class of causal explanations about the relationship between a sense of belonging and academic outcomes. Along these lines, Bryk et al. (2013) describes an intervention for African American community college students in developmental mathematics courses, a field where they are stereotyped as not belonging. It uses a large, tailored menu of activities to assure a sense of belonging and has

seen positive outcomes. One might therefore be reassured that diverse belonging interventions are likely effective, so long as they are strong enough to achieve the intermediate outcomes: for example, providing enough signals of belonging to produce a sense of belonging. More generally, we reduce uncertainty about generalization when diverse activities, all of which are consistent with a core component and theoretical construct, are found to achieve specified intermediate outcomes. Reflective practitioners can often contribute to such diverse activities, just as they did for HIV prevention.

Can Interactions Between Theory and Practice be Made More Systematic?

Fitting the Theoretical Constructs to New Contexts

In ongoing collaboration between RWJF and the CDC, we have observed several ways that researchers and developers work on adaptation when they are confronted with the challenges and opportunities of new contexts.

- The CDC developed detailed adaptation packages for evidence-based HIV prevention in its Diffusion of Evidence-Based Interventions (DEBI; https://effectiveinterventions.cdc.gov/en/HighImpactPrevention/Interventions.aspx). DEBI follows a relatively painstaking process of assessing the characteristics of new populations, organizations, and environments, then considering individual core components in light of these, then indicating what is permitted, not permitted, and questionable. These adaptation packages are being emulated for evidence-based violence prevention programs (e.g., Morrel-Samuels et al., 2014).
- Rotheram-Borus and colleagues distill the essential components of evidence-based practice, then train and supervise until practitioners have a flexible grasp of ways to use the essential components (Chorpita et al., 2007; Rotheram-Borus, Swendeman, & Chorpita, 2012).
- Hawe, Shiell, and Riley (2004) consider the function that a component is supposed to achieve, along the lines of "many roads" leading to intermediate objectives.
- Quality improvement collaboratives in medicine (Institute for Healthcare Improvement, 2015) and education (Bryk et al., 2013) utilize principles from industry that encourage practitioners to try rapid-cycle adaptations to achieve an improvement aim. Evidence

for their impact is positive but limited (Schouten, Hulscher, van Everdingen, Huijsman, & Grol, 2008).

- Some developers consider new contexts and "reinvent" the entire program for a range of new situations, notably the *Positive Parenting Program* (Leung, Sanders, Leung, Mak, & Lau, 2003; Nowak & Heinrichs, 2008; Prinz, Sanders, Shapiro, Whitaker, & Lutzker, 2009; Whittingham, Sofronoff, Sheffield, & Sanders, 2008) and a family of programs to prevent teen dating violence (Foshee et al., 2004, 2012, 2015).
- Recognizing that resources, time, and other context issues may interfere with implementation, the Multiphase Optimization Strategy (Collins, 2013) develops program theory, pares away irrelevancies, then tests the optimal program components, often through factorial randomized experiments, and evaluates the optimized intervention against a suitable comparison condition.
- Still others deal with the reality that practitioners tend to adopt individual components, rather than entire models. For example, Embry and Biglan (2008) suggest that practitioners incorporate individual evidence-based components if they cannot or will not adopt entire models.

Toward a Better Use of Practitioner Knowledge

Each of these approaches has merit, but relatively few of them take advantage of what reflective practitioners know. From Table 1, we believe there is a way to maximize the strengths of both researcher–developers and practitioners to probe external validity and in particular, the better understanding of prevalent and important treatment variations. RWJF is supporting a variety of pilot projects to develop efficient ways to characterize and assess local adaptations. Of course, we are in favor of additional, rigorous tests of adaptation, which ideally would follow some ways to characterize and quantify local adaptations. However, given the range of potential adaptations, it is necessary to prepare—in Collins' (2013) terms to optimize—by setting priority on those adaptations that will do the most to reduce uncertainty for the decision maker (Cronbach & Shapiro, 1982).

If they were collected more systematically, practitioner experiences could become a resource for external validity, program model refinement, and the assessment of local adaptation. Some evaluation approaches do in fact consult practitioners for theory building (e.g., Chen, 2010), which gains important information as well as buy-in. However, they don't necessarily

distinguish reflective practitioners from novices. Communities of practice offer another vehicle to aggregate such experience and develop a shared repertoire of resources (Wenger, McDermott, & Snyder, 2002). We have observed that practitioners are often hungry to share these challenges and opportunities.

Capturing a Higher Volume of Contexts

A mechanism is needed to aggregate information about contexts and the adaptations of interventions that occur in those contexts. Doing so helps us to see the patterns in context, so that we can make informed choices about what to study and how to exercise both quality control and quality improvement. By consulting diverse practitioners in the process, one can overcome the potential problem of bias in the settings and populations they have experienced. One potential model comes from quality improvement in medicine (Øvretveit, Leviton, & Parry, 2011). Adaptations at many sites, achieved through rapid cycle improvements, are reported, refined, assessed, and shared. Together they aim to further refine theory. Another possibility is to use ''crowdsourcing'' to amass such information, either through Internet interactions or possibly a survey. A third possibility is to engage technical assistance providers or reflective practitioners of the EBIs, since they will have seen a fairly wide variety of contexts. The CDC is piloting crowdsourcing and debriefing technical assistance providers for violence prevention projects, under a grant from RWJF.

Focusing the Interaction of Practice and Theory

The high volume of contexts would then be winnowed and focused on adaptations that may be worthy of further study. We believe there are two key principles to do so. First, with Hill, Maucione, and Hood (2007), we agree that assessment should be focused on adaptations that occur frequently or on the barriers to implementation that occur frequently. Assessing the frequently encountered adaptations and barriers will do the most to effect population impact, by assisting widespread, high-quality implementation. Two projects are assessing the frequency of adaptations and barriers, through surveys, key informant interviews, and focus groups.

Second, the examination of core components and program theory allows us to assess which barriers and facilitators have powerful effects on implementation and also which adaptations are most important, in that they are likely to be helpful, harmful, irrelevant, or controversial. Irrelevancies may

appear benign, but they may still be undesirable when they cost time and money (Collins, 2013). Initially, an interaction of researchers and practitioners might assess adaptation using informed judgment (Backer, 2001). This may be sufficient in many cases, given the logic of induction for ruling out and for assessing strength and integrity. In other cases, rigorous tests might be needed of the adaptations' effects on mediators, moderators, intermediate, or ultimate outcomes. By focusing, however, we can determine where the greatest uncertainty lies for decision makers. It may well be that tests of the controversial adaptations take the highest priority for their information value in reducing uncertainty (Cronbach & Shapiro, 1982). Several projects are currently assessing this approach.

Conclusions

Assessment of external validity is a marathon, not a sprint; it requires a program of inquiry and a body of evidence from diverse sources. Given the small samples of sites and other difficulties in statistical methods for generalization, we have what Tipton et al. (**In Press**) have rightly called "a missing data problem." Under these circumstances, we would recommend, not just mixed method studies, but mixed method programs of inquiry to better understand context and give the quantitative evaluators the tools they need for generalization.

Program theory is essential to an adequate assessment of external validity in general and to the assessment of local adaptation, in particular. Interaction between the researcher–developer and practitioners has potential to improve these assessments. Ideally, the process could become iterative (Øvretveit et al., 2011), using ex ante theory to test adaptations selectively, then using the results ex post to further sharpen the program theory and extend generalizations. Core components are most usefully conceived as a family of related activities all aiming at the same intermediate outcome, mediator, or moderator of effects. Local adaptations can be inductively ruled out or provisionally ruled in, based on core components' presence, absence, and apparent strength. Initially, this process depends on judgment, but selective tests would ideally follow, prioritized for their value to reduce uncertainty.

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.

References

Avellar, S. A., Kleinman, R., Sama-Miller, E., et al. (**In Press**). External validity: The next step for systematic reviews? *Evaluation Review*.

Babitsch, B., Gohl, D., & von Lengerke, T. (2012). Re-revisiting Andersen's behavioral model of health services use: A systematic review of studies from 1998–2011. *GMS Psycho-Social-Medicine*, *9*. doi:http://doi.org/10.3205/psm000089

Backer, T. E. (2001). *Finding the balance: Program fidelity and adaptation in substance abuse prevention: A state of-the-art review*. Retrieved from http://modelprograms.samhsa.gov/pdfs/Finding-Balance1.pdf

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*, 497–529.

Berwick, D. (2003). Disseminating innovation in health care. *Journal of the American Medical Association*, *289*, 1969–1975.

Bishop, D. C., Pankratz, M. M., Hansen, W. B., Albritton, J., Albritton, L., & Strack, J. (2014). Measuring fidelity and adaptation reliability of an instrument for school-based prevention programs. *Evaluation & the Health Professions*, *37*, 231–257.

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. S., & Roitman, D. B. (1987). The fidelity adaptation debate: Implications for the implementation of public sector social programs. *American Journal of Community Psychology*, *15*, 253–268.

Bryk, A. S., Yeager, D. S., Hausman, H., Muhich, J., Dolle, J. R., Grunow, A., & Gomez, L. (2013, June). *Improvement research carried out through networked communities: Accelerating learning about practices that support more productive student mindsets*. A white paper prepared for the White House meeting on "Excellence in Education: The Importance of Academic Mindsets." Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from http://www.carnegiefoundation.org/resources/publications/improvement-research-carried-networked-communities-accelerating-learning-practices-support-productive-student-mindsets/

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Canadian Institutes of Health Research. (2014). *About knowledge translation*. Ottawa, Canada. Retrieved from http://www.cihr-irsc.gc.ca/e/29418.html

Castro, F. G., Barrera, M., Jr., & Holleran Steiker, L. K. (2010). Issues and challenges in the design of culturally adapted evidence-based interventions. *Annual Review of Clinical Psychology*, *6*, 213–239.

Center for Mental Health in Schools at University of California, Los Angeles. (2008). *Conduct and behavior problems: Intervention and resources for school aged youth*. Los Angeles, CA: Author. Retrieved from http://smhp.psych.ucla.edu

Centers for Disease Control and Prevention. (2014). *Compendium of evidence-based interventions and best practices for HIV prevention*. Retrieved from http://www.cdc.gov/hiv/prevention/research/compendium/index.html

Chang, H. (2012, October 11). *Scientific pluralism and the mission of history and philosophy of science*. [YouTube] Inaugural lecture, University of Cambridge. Retrieved from www.youtube.com/watch?v=zGUsIf9qYw8

Chen, H. T. (2010). The bottom-up approach to integrative validity: A new perspective for program evaluation. *Evaluation and Program Planning*, *33*, 205–214.

Chen, H. T., & Rossi, P. H. (1992). *Using theory to improve program and policy evaluations*. Westport, CT: Greenwood.

Chorpita, B., Becker, K., & Daleiden, E. (2007). Understanding the common elements of evidence-based practice: Misconceptions and clinical examples. *Journal of the American Academy of Child and Adolescent Psychiatry*, *46*, 647–652.

Collins, L. M. (2013). Optimizing family interventions: The multiphase optimization strategy (MOST). In S. McHale, P. McHale, P. Amato, & A. Booth (Eds.), *Emerging methods in family research* (pp. 231–244). New York, NY: Springer.

Copi, I. M., Cohen, C., & Flage, D. E. (2007). *Essentials of logic* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Coyle, S. L., Needle, R. H., & Normand, J. (1998). Outreach-based HIV prevention for injecting drug users: A review of published outcome data. *Public Health Reports*, *113*, 19–30.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Cronbach, L. J., & Shapiro, K. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Davidoff, F., Dixon-Woods, M., Leviton, L., & Michie, S. (2015). Demystifying theory and its use in improvement. *BMJ Quality and Safety*, *24*, 228–238. doi:10.1136/bmjqs-2014-003627

Del Grosso, P., Kleinman, R., Mraz Esposito, A., Sama-Miller, E., & Paulsell, D. (2014). *Assessing the evidence of effectiveness of home visiting program models implemented in tribal communities*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Dreyfus, H. L., & Dreyfus, S. E. (1988). *Mind over machine: The power of human intuition and expertise in the era of the computer*. New York, NY: Free Press.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–350.

Dusenbury, L., Brannigan, R., Hansen, W. B., Walsh, J., & Falco, M. (2005). Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research*, *20*, 308–313.

Eccles, M. P., & Mittman, B. S. (2006). Welcome to Implementation Science. *Implementation Science*, *1*. doi:10.1186/1748-5908-1 -1

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

Embry, D. D., & Biglan, A. (2008). Evidence-based kernels: Fundamental units of behavioral influence. *Clinical Child and Family Psychological Review*, *11*, 75–113.

Emshoff, J. G., Blakely, C., Gottschalk, R., Mayer, J., Davidson, W. S., & Erickson, S. (1987). Innovation in education and criminal justice: Measuring fidelity of implementation and program effectiveness. *Educational Evaluation and Policy Analysis*, *9*, 300–311.

Epstein, D., & Klerman, J. A. (2013). When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model. *Evaluation Review*, *36*, 375–401.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The Reasoned Action approach*. New York, NY: Psychology Press.

Fixsen, D. (2015). *Our approach. National Implementation Research Network*. Retrieved from http://nirn.fpg.unc.edu/about-nirn/our-approach

Flay, B. R., Berkowitz, M. W., Bier, M. C., & The Social and Character Development Research Consortium. (2009). Elementary school-based programs theorized to support social development, prevent violence, and promote positive school climate: Description and hypothesized mechanisms of change. *Journal of Research in Character Education*, *7*, 21–50.

Foshee, V. A., Bauman, K. E., Ennett, S. T., Linder, G. F., Benefield, T. S., & Suchindran, C. (2004). Assessing the long-term effects of the Safe Dates program and a booster in preventing and reducing adolescent dating violence victimization and perpetration. *American Journal of Public Health*, *94*, 619–624.

Foshee, V. A., Dixon, K. S., Ennett, S. T., Moracco, K. E., Bowling, J. M., Chang, L., & Moss, J. L. (2015). The process of adapting a universal dating abuse prevention program to adolescents exposed to domestic violence. *Journal of Interpersonal Violence*, *30*, 2151–2173.

Foshee, V. A., Reyes, H. L. M., Ennett, S. T., Cance, J. D., Bauman, K. E., & Bowling, J. B. (2012). Assessing the effects of Families for Safe Dates, a

family-based teen dating abuse prevention program. *Journal of Adolescent Health*, *51*, 349–356.

Freire, K. E., Perkinson, L., Morrel-Samuels, S., & Zimmerman, M. A. (2015). Three Cs of translating evidence-based programs for youth and families to practice settings. *New Directions for Child and Adolescent Development*, *2015*, 25–39.

Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: The RE-AIM framework. *American Journal of Public Health*, *89*, 1323–1327.

Griner, D., & Smith, T. B. (2006). Culturally adapted mental health interventions: A metaanalytic review. *Psychotherapy: Theory, Research, Practice, Training*, *43*, 531–548.

Hawe, P., Shiell, A., & Riley, T. (2004). Complex interventions: how "out of control" can a randomised controlled trial be? *British Medical Journal*, *328*, 1561–1563.

Hausmann, L. R., Schofield, J. W., & Woods, R. L. (2007). Sense of belonging as a predictor of intentions to persist among African American and White first-year college students. *Research in Higher Education*, *48*, 803–839.

Hill, L. G., Maucione, K., & Hood, B. K. (2007). A focused approach to assessing program fidelity. *Prevention Science*, *8*, 25–34.

Hoffmann, T., Glasziou, P., Boutron, I., Milne, R., Perera, R., Moher, D., … Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *British Medical Journal*, *348*, g1687.

Horne, C. S. (**In Press**). Assessing and strengthening evidence-based program registries' usefulness for social service program replication and adaptation. *Evaluation Review*.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion*. New Haven, CT: Yale University Press.

Hulleman, C., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research Educational Effectiveness*, *2*, 88–110.

Institute for Healthcare Improvement. (2015). *About us: The science of improvement*. Retrieved from http://www.ihi.org/about/Pages/ScienceofImprovement.aspx

International Initiative for Impact Evaluation. (2015). *Evidence matters*. Retrieved from http://blogs.3ieimpact.org/evidence-matters/

Jagers, R. J., Sydnor, K., Mouttapa, M., & Flay, B. R. (2007). Protective factors associated with preadolescent violence: Preliminary work on a cultural model. *American Journal of Community Psychology*, *40*, 138–145.

Johnson, K. (2009). *State-based home visiting: Strengthening programs through state leadership*. New York, NY: National Center for Children in Poverty. Retrieved from http://www.nccp.org/publications/pdf/text_862.pdf

Lattimore, D., Griffin, S. F., Wilcox, S., Rheaume, C., Dowdy, D., Leviton, L. C., & Ory, M. G. (2010). Understanding the challenges and adaptations made by community organizations for translation of evidence-based behavior change physical activity interventions: A qualitative study. *American Journal of Health Promotion*, *24*, 427–434.

Leung, C., Sanders, M. R., Leung, S., Mak, R., & Lau, J. (2003). An outcome evaluation of the implementation of the Triple P—Positive Parenting Program in Hong Kong. *Family Process*, *42*, 531–544.

Leviton, L. C. (2015). Evaluation practice and theory: Up and down the ladder of abstraction. *American Journal of Evaluation*, *36*, 238–242.

Leviton, L. C., & Guinan, M. E. (2003). HIV prevention and the evaluation of public health programs. In R. O. Valdiserri (Ed.), *Dawning answers: How the HIV/AIDS epidemic has helped to strengthen public health* (pp. 155–176). Oxford, England: Oxford University Press.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. *New Directions for Program Evaluation*, *1993*, 5–38.

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders*, *4*, 124–147.

March, J. G. (1994). *A primer on decision making: How decisions happen*. New York, NY: Free Press.

Michie, S., West, R., Campbell, R., Brown, J., & Gainforth, H. (2014). *An ABC of behavior change theories*. London, England: Silverback.

Morrel-Samuels, S., Hutchison, P., Perkinson, L., Bostic, B., & Zimmerman, M. (2014). *Selecting, implementing and adapting Youth Empowerment Solutions*. Ann Arbor: University of Michigan.

Naylor, M. D., Feldman, P. H., Keating, S., Koren, M. J., Kurtzman, E. T., Maccoy, M. C., & Krakauer, R. (2009). Translating research into practice: Transitional care for older adults. *Journal of Evaluation in Clinical Practice*, *15*, 1164–1170.

Nowak, C., & Heinrichs, N. (2008). A comprehensive meta-analysis of Triple P—Positive Parenting Program using hierarchical linear modeling: Effectiveness and moderating variables. *Clinical Child and Family Psychology Review*, *11*, 114–144.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge, England: Cambridge University Press.

Øvretveit, J., Leviton, L. C., & Parry, G. (2011). Increasing the generalisability of improvement research with an improvement replication programme. *BMJ Quality and Safety*, *20*, i87–i91.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York, NY: Guilford.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.

Peters, R. H., & Wexler, H. K. (2005). *Substance abuse treatment for adults in the criminal justice system: A treatment improvement protocol* (DHHS Publication No. (SMA) 05-4056). Rockville, MD: Substance Abuse and Mental Health Services Administration.

Perkinson, L. (2012). *Environmental scan of adaptation guidance* (Unpublished report). Atlanta, GA: CDC Foundation.

Pressman, J. L., & Wildavsky, A. B. (1984). *Implementation*. Berkeley: University of California Press.

Prinz, R. J., Sanders, M. R., Shapiro, C. J., Whitaker, D. J., & Lutzker, J. R. (2009). Population-based prevention of child maltreatment: The U.S. Triple P system population trial. *Prevention Science*, *10*, 1–12.

Prochaska, J. O., & Norcross, J. C. (2009). *Systems of psychotherapy: A transtheoretical analysis* (7th ed.). Belmont, CA: Wadsworth.

Ringwalt, C. L., Ennett, S., Johnson, R., Rohrbach, L. A., Simons-Rudolph, A., & Vincus, A. (2003). Factors associated with fidelity to substance use prevention curriculum guides in nation's middle schools. *Health Education & Behavior*, *30*, 375–391.

Robinson, T. N., Matheson, D. M., Kraemer, H. C., Wilson, D. M., Obarzanek, E., Thompson, N. S., … Killen, J. D. (2010). A randomized controlled trial of culturally-tailored dance and reducing screen time to prevent weight gain in low-income African-American girls: Stanford GEMS. *Archives of Pediatrics & Adolescent Medicine*, *164*, 995–1004.

Rogers, E. (2003). *Diffusion of innovations* (5th ed.). New York, NY: Free Press.

Rohrbach, L. A., Grana, R., Sussman, S., & Valente, T. W. (2006). Type II translation: Transporting prevention interventions from research to real-world settings. *Evaluation and the Health Professions*, *29*, 302–333.

Rossi, P. H. (1987). The Iron Law of evaluation and other metallic rules. *Research in Social Problems and Public Policy*, *4*, 3–20.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Rotheram-Borus, M. J., Swendeman, D., & Becker, K. D. (2014). Adapting evidence-based interventions using a common theory, practices and principles. *Journal of Clinical Child and Adolescent Psychology*, *43*, 229–243.

Rotheram-Borus, M. J., Swendeman, D., & Chorpita, B. (2012). Disruptive innovations for designing and diffusing evidence-based interventions. *American Psychologist*, *67*, 463–476.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York, NY: Basic Books.

Schouten, L. M., Hulscher, M. E., van Everdingen, J. J., Huijsman, R., & Grol, R. P. (2008). Evidence for the impact of quality improvement collaboratives: Systematic review. *British Medical Journal*, *336*, 1491–1494.

Schuh, R. G., & Leviton, L. C. (2006). A framework to assess the development and capacity of nonprofit agencies. *Evaluation and Program Planning*, *29*, 171–179.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theorists and their theories*. Newbury Park, CA: Sage.

Stake, R. E., & Schwandt, T. A. (2006). On discerning quality in evaluation. In I. Shaw, J. C. Greene, & M. Mark (Eds.), *The SAGE handbook of evaluation* (pp. 404–418). Thousand Oaks, CA: Sage.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology*, *34*, 379–440.

Stuart, E. A., & Rhodes, A. (**In Press**). Assessing the external validity of randomized trial results: A case study in the difficulties of finding sufficient data. *Evaluation Review*.

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (**In Press**). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*.

Tipton, E., & Peck, L.R. (**In Press**). A design-based approach to improve external validity in welfare policy evaluations. *Evaluation Review*.

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*, 82–96.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, *331*, 1447–1451.

Walton, G. M., Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2012). *A brief intervention to buttress middle school students' sense of social-belonging: Effects by race and gender*. Unpublished manuscript, Stanford University, Palo Alto, CA.

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *Journal of Educational Psychology*, *107*, 468.

Weiss, C. H. (1997). *Evaluation: Methods for studying programs and policies* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Wenger, E., McDermott, R., & Snyder, W. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Cambridge, MA: Harvard Business School Press.

Whittingham, K., Sofronoff, K., Sheffield, J., & Sanders, M. R. (2008). Stepping Stones Triple P: An RCT of a parenting program with parents of a child

diagnosed with an Autism Spectrum Disorder. *Journal of Abnormal Child Psychology*, *37*, 469–480.

Wilcox, S., Dowda, M., Leviton, L. C., Bartlett-Prescott, J., Bazzarre, T., Campbell-Voytal, K., . . . Wegley, S., et al. (2008). The active for life initiative: Final results of translating two evidence-based physical activity programs into practice. *American Journal of Preventive Medicine*, *35*, 340–351.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods*, *6*, 413–429.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, *49*, 156–167.

Author Biographies

Laura C. Leviton, PhD, is the senior adviser for evaluation at the Robert Wood Johnson Foundation in Princeton, New Jersey, a position that the Foundation created for her to advise and consult on over 100 evaluations across its many initiatives and national programs. She is coauthor with William Shadish and Thomas Cook of *Foundations of Program Evaluation* (Sage, 1991) and received several national awards, including the American Evaluation Association (AEA) Outstanding Publication Award for *The Systematic Screening and Assessment Method: Finding Innovations Worth Evaluating. New Directions in Evaluation*, 2010 (125). She was president of AEA and served on three Institute of Medicine committees, among other national and international positions.

Mathew D. Trujillo, PhD, is a research associate in the Research, Evaluation and Learning unit of the Robert Wood Johnson Foundation. In 2012, he received a doctorate in psychology and social policy from the Woodrow Wilson School of Public and International Affairs at Princeton University. Previously, he served an adjunct researcher with the RAND Corporation.