Examining Real-World Applicability of Depression Prevention Trials

By

Jennifer Marie Stewart

Thesis
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
in
Psychology
August 2015
Nashville, Tennessee

Approved:

Judy Garber, Ph.D.
Steven Hollon, Ph.D.

ACKNOWLEDGEMENTS

I am grateful to all of those with whom I have had the pleasure to work on this, as well as other projects during my time at Vanderbilt. I am especially appreciative of Dr. Steven Brunwasser for his tremendous mentorship on this project and his helping me to grow as a researcher. Finally, I would like to thank my advisor, Dr. Judy Garber, who has supported me through this project and hopefully through many more.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Prevention research has been burgeoning over the past decade. Evidence of the efficacy of preventive interventions exists for a variety of conditions such as depression, violence, tobacco, alcohol, and substance use, and academic failure (Clarke et al., 1995; Flannery et al., 2003; Biglan et al., 2004; Botvin et al., 1995; Gunn et al., 2002). Although the field of prevention science is progressing, deciding when a program is ready for dissemination remains a challenge. The Society for Prevention Research (SPR) developed (Flay et al., 2005) and recently updated (Gottfredson et al., 2015) a set of standards to assist practitioners, policy makers, and administrators determine whether an intervention is efficacious, effective, or ready for dissemination. According to these standards, efficacy trials are conducted under optimally controlled conditions with the goal of producing clinically meaningful effects. Once efficacy has been established, effectiveness trials are conducted to demonstrate that significant effects can be found under more real-world conditions. Finally, when the standards for efficacy and effectiveness have been met, programs then can be evaluated for dissemination.

A common view is that as programs move from tightly controlled trials to more real-world contexts, the effects of the intervention diminish. This drop in effects has occurred for programs that previously were found to be efficacious, but show negligible effects in real-world settings (Chambers, Glasgow, & Stange, 2013; Weisz & Jensen, 2001). For example, studies have found psychotherapy to have stronger effects when delivered in university-based settings than in community clinics (Weisz, Dononberg, Han & Weiss, 1995). A recent meta-analysis of depression prevention programs for youth found that whereas several prevention programs have

been promising in terms of efficacy, none has demonstrated sufficient evidence of effectiveness under realistic conditions (Brunwasser & Garber, 2015), which likely is due to moving from tightly controlled to more real-world contexts. To our knowledge, however, data do not exist that explicitly confirm this hypothesis, or that identify which specific components of a program contribute to this decline. Factors such as the providers, the organization or setting, the amount of training, loss in fidelity, or the participants themselves may be driving this drop in effects, but this issue has not been adequately addressed empirically (Gillham et al., 2006).

By examining the research cycle dimensionally as studies move from efficacy to effectiveness, we can test whether programs with more realistic delivery truly do show diminishing effects over time, and which components of these programs are driving this decline. The current paper describes the operationalization of this process and development of a rating scale for quantifying the level of real-world applicability of a program.

Background

The SPR Standards of Evidence provide guidelines for determining whether prevention programs have demonstrated adequate evidence of efficacy and effectiveness to justify widespread implementation (Flay et al., 2005). Recently, these standards of evidence were updated to reflect changes in the field of prevention science and to promote greater flexibility in the research cycle (Gottfredson et al., 2015). Several research teams also have developed criteria for evaluating evidence-based interventions and for determining when a program is efficacious, effective, or ready for dissemination (Gartlehener et al., 2006; Glasgow, Vogt, & Boles, 1999; Kocsis et al., 2010; Mason et al., 2013; Spoth, et al., 2013; Wandersman et al., 2008).

Using effectiveness and efficacy standards to guide program development, however, is not sufficient for predicting which programs will be successful in real-world settings. Rather than being purely pragmatic or explanatory, trials can display varying degrees of effectiveness (Thorpe et al., 2001). Adopting a dimensional perspective, however, does not necessarily rule out the utility of categorizing trials as efficacious and effective. Conceptualizing the research cycle as moving from efficacy to effectiveness to dissemination in a categorical way can guide research questions and inform stakeholders about the progress and development of an intervention. Nevertheless, if the goal is to determine which aspects of an intervention drive the effects, or alternatively, contribute to null effects, then a dimensional approach to evaluating programs may be preferred.

The PRECIS framework (Thorpe et al., 2001) uses a dimensional system to assess the degree to which an intervention study aligns with its stated purpose as an efficacy or effectiveness trial. The PRECIS system rates clinical trials regarding the extent to which they are pragmatic or explanatory, and the degree to which they are implemented under optimal vs. realistic conditions (i.e., as an efficacy or effectiveness trial). Stakeholders then can evaluate how much these components of a trial actually align with their research goals (i.e., did the research questions match the level of pragmatic vs. explanatory components) (Winter & Colditz, 2014).

PRECIS carefully qualifies studies as efficacy or effectiveness trials, which can guide program evaluation and selection. However, it does not allow us to determine the extent to which intervention delivery and contextual factors were related to the magnitude of program effects. Often trials are labeled efficacy or effectiveness based on a single study characteristic (e.g., provider, setting), but many factors determine the extent to which a trial is delivered under real-

world conditions. An important aim of clinical trials is to understand *why* programs succeed; that is, what factors contribute to success (Weisz & Jenson, 2001).

With this goal in mind, we developed a rating system designed to capture the degree to which trials reflect optimal research conditions (efficacy trials) vs. real-world conditions (effectiveness trials) using a dimensional scale. Similar to the PRECIS criteria, we conceptualized studies as being on a continuum from efficacy to effectiveness. Unlike the PRECIS program, however, our scale does not delineate between effectiveness and efficacy; rather, the measure allows researchers to explore a program's level of realism dimensionally, thereby informing program development and the analysis of possible moderating factors.

We developed a rating scale with the following goals in mind:

1. To identify factors common across studies that deal with the components of real-world applicability (i.e., ecological validity and external validity).
2. To evaluate research trials regarding the extent to which they report these various factors.
3. To use these factors to evaluate the hypothesis that there is a negative relation between the level of realistic conditions and sustained intervention effects.
4. To identify which of the real-world factors predict to the magnitude of the effect sizes.

CHAPTER II

SCALE DEVELOPMENT

Real-World Applicability Framework

In developing this scale, we first identified the components that contribute to an intervention being successful in a real-world setting, and then disassemble those into their own factors. *Figure 1* presents the conceptual basis of the real-world applicability framework. Internal validity, external validity, ecological validity, and transportability all contribute to an intervention being successful in a real-world setting. We examined the specific aspects of each of these domains to determine the characteristics most (or least) related to a program's sustained effects.

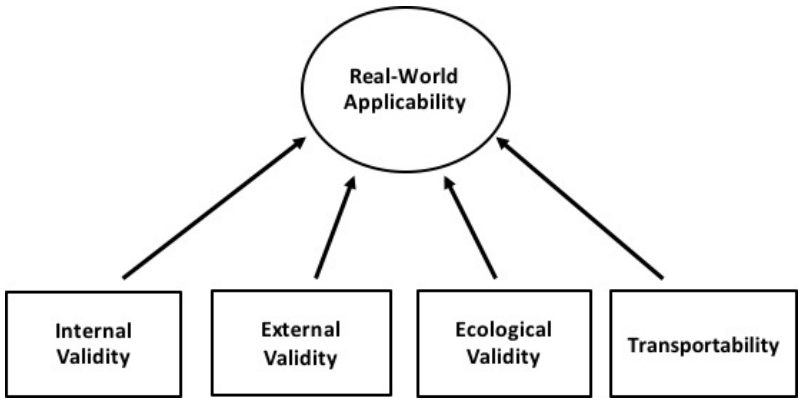

Figure 1. Real-World Applicability Framework

Most criteria developed to evaluate prevention trials have focused on capturing broad components found universally across multiple fields. For example, the Standards of Evidence include criteria that apply to programs that target varying populations, study durations,

diagnoses, etc., making it relevant to the majority of prevention trials. Being so broad, however, compromises specificity and the ability to capture and account for small design differences between studies. As depression prevention researchers, we included prototypical examples from depression prevention trials as anchor points for most of the criterion in the scale. Nevertheless, the framework and many of the scale criteria used here likely are applicable to clinical trials across prevention trials and outside the depression prevention literature.

The Real-World Applicability scale includes six items related to external validity and eleven items about ecological validity. We focused here on external and ecological validity because program characteristics within these domains are especially relevant to moving successfully from "ideal" to "real" intervention settings. Moreover, because tools already exist that evaluate internal validity (Higgins et al., 2011), we did not make that a focus here. Using existing measures of internal validity in conjunction with our measure of external and ecological validity would allow researchers to determine the overall quality *and* real-world applicability of a particular program.

Scoring

Each item on the scale is coded on a 1-5-point scale. We used a 5- as compared to a 3-point response scale in order to capture as much variability in program characteristics as possible. Items can be coded as half-points between anchors (e.g., 2.5, 1.5) when appropriate. In cases where the information necessary to assign a score is missing, an arbitrary code of -999 is used. We selected items on the scale specifically to reflect characteristics of depression prevention trials. The response anchors for each item describe what most often happens in trials, as well as features that we would make the trials closer to real-world practice.

Coding of programs on this scale requires using some judgment. We do not assume that different coders will reach a perfect consensus, but we do expect consistency in scores. The ultimate goal is to have a correspondence in the rank order of studies among coders, rather than matching on exact scores. We constructed the scale to allow coders to distinguish between studies that are higher and lower in real-world applicability, and to do so in a consistent way.

Domain and Item Descriptions

Before using this scale to examine an intervention program, it is important to define the target population of interest, with regard to characteristics such as age, demographics, risk factors, etc., in order to clarify to whom the intervention's effects should generalize. Some studies do not adequately specify which individuals are the presumed target of the intervention. For example, if a school-based intervention is delivered after school to youths in 7th grade, it may be unclear if this was the intended plan, or if the program was actually meant to be provided during school to all middle school students. Therefore, investigators should state explicitly whether the intervention sample and settings were intentional or simply convenient. Such information is essential in order to identify what aspects of these programs contribute to the observed effects, and how to adjust these programs for more realistic delivery.

Internal Validity

Internal validity refers to the extent to which a study is free from bias and the statistical inferences are valid (Higgins et al., 2001). In general, internal validity precedes concerns about whether a program can be effective in a real-world setting. It may be necessary, however, to sacrifice some internal validity when moving toward effectiveness. For example, randomizing by classrooms or schools when delivering a school-based program may compromise some internal

validity, but likely is closer to how the program eventually would be implemented in the real word.

In the development of the current scale, we focused on external and ecological validity, although we recognize the importance of internal validity when assessing real-world applicability. Fortunately, other methods exist that assess internal validity. Use of other methods for assessing internal validity such as the Cochran Collaboration's tool in combination with the current scale may facilitate the identification of programs that maintain high levels of internal validity as they move towards more realistic implementation.

Transportability

Transportability is another domain that is important when evaluating a program's level of real-world applicability. Transportability refers to the extent to which researchers can engage with and apply an intervention in realistic contexts. Although, criteria for coding transportability are not included in the current scale, future adaptations of the scale will include items related to this domain. Examples of relevant program characteristics are multiplatform capabilities (e.g., computer-based; in-person), the intensity of supervision and training of providers, and the "dosage" of the program (e.g., number of sessions; length of sessions).

External Validity

External validity is concerned with the representativeness of the sample; that is, external validity refers to the extent to which inferences based on the sample are relevant to the population of interest. As programs move towards effectiveness, external validity captures the generalizability of their effects to the targeted population (Rothwell, 2005). Several items within our scale (e.g., Extrinsic Incentives, Inclusion/Exclusion Criteria) may be relevant to both the

external and ecological validity of a program. Nevertheless, we grouped items according to whether they contributed most to participant/sample representativeness (i.e., external validity) or design/contextual representativeness (i.e., ecologically validity). The following items are included in our scale to assess the level of external validity of a program:

1. *Intervention Setting*

This item examines the extent to which there is access to the target population in the intervention setting. Sampling from multiple settings and including sites with high access to the target population would indicate lower threat to external validity than sampling from single sites and settings with limited access to the targeted population. Excellent access would indicate that the vast majority of the target population is accessible. This would most likely be through sampling multiple different settings, such as schools or pediatrician offices. At the other end of the spectrum would be settings that offer highly limited access to the population of interest. For example, a universal prevention study that only recruited participants from a boy's basketball camp would have low access to the population of interest.

2. *Breadth of Primary Sampling Units (PSU)*

This item measures the degree to which primary sampling units (PSUs; e.g., schools, clinics, neighborhoods) were selected in a manner that optimized generalizability to the population of PSUs. The term PSU describes the setting(s) from which investigators recruit the study sample. This item focuses on the representativeness of the PSUs to the population of interest. That is, how broad was the initial sample from which eligible PSUs were recruited. The scoring on this item ranges from a very narrow focus at the low end, such as using only a single

PSU in a confined area (e.g., a single school, neighborhood, or clinic), to the highest score for using multiple PSUs in multiple regions or areas.

3. *Representativeness of those Screened*

Scores on this item are based on the approximate percentage of individuals within the PSUs that were assessed for study inclusion criteria. For targeted interventions that require screening to determine study eligibility, selection bias may be present when potentially eligible individuals do not complete the screening procedure and their eligibility for the study is never determined. Studies that do not require a screening process (i.e., all individuals in the PSUs are eligible), are given the highest score on this item. For studies that include multiple stages of screening, this item considers the combined total percentage of participants assessed at *each* stage of screening.

4. *Enrollment Rate for Individuals Meeting Inclusion Criteria & Invited to Participate*

This criterion reflects what percentage of individuals invited to participate actually enrolled in the study. Not all individuals who meet the study inclusion criteria and are offered spots in the study end up participating, which creates a threat to external validity. For universal studies, we calculate this score by dividing the number of participants randomized by the number of total participants in the PSUs. In targeted studies, we calculate scores on this item by dividing the number of participants randomized by the number of individuals offered spots in the study.

5. *Self-Selection Bias*

This item evaluates the extent to which randomized participants differ from those invited to participate, but did not (i.e., decliners). Coders rate this item based on the number of variables

that differ between the randomized participants, and those who were eligible but declined, as well as the degree of importance of the variable. For example, coders rate differences in outcome variables, key predictors, or hypothesized mediators lower than differences on only demographic variables. Studies often do not report this item, as individuals who choose not to participate are not typically included in analyses. However, this creates a threat to external validity. In cases where there are no decliners or the percentage of decliners is less than 5%, a score of *5* is given.

Ecological Validity

The criteria relating to ecological validity refer to the extent to which conditions of intervention delivery in the study resemble conditions expected if delivered in the real world. Originally, the term ecological validity indicated the degree of correlation between a proximal cue and the distal object to which it was related (Brunswik, 1956; Hammond & Brunswik, 1998). However, the concept of ecological validity has been adapted to reflect design representativeness within a study. Despite some researchers arguing that use of the term ecological validity to reflect contextual or design representativeness is incorrect, this definition is most common in the field and is the one upon which we based this scale. The following items are included in our scale to assess the level of ecological validity of a program:

1. *Provider Accessibility*

This item captures the extent to which the intervention providers were accessible in the intervention setting. Delivery of a program in a realistic setting ideally should be provided by facilitators who are naturally present in that environment (e.g., teachers, community clinicians, nurses), or should require no provider, such as with web-based or self-paced programs. In many research studies, however, members of the investigative team or the actual program developers

often implement the intervention. Coders would rate such studies low on "realism" regarding provider accessibility.

2. *Accessibility of Provider Trainers*

It is also important to specify who is responsible for training providers, and their level of accessibility. This item assesses the extent to which the intervention providers needed to be trained, and how accessible the trainers were to them. In many studies, program developers or members of the research team are responsible for training study therapists. These expert trainers, however, are not often available or accessible in the actual intervention settings. Having trainers that already are part of these environments (e.g., teachers, school counselors, or community clinicians) is more realistic. Coders give the highest scores on this item when training is not necessary or if providers can self-train or train online.

3. *Accessibility of Provider Supervision*

This item evaluates the extent to which the providers received supervision from sources that likely are not readily accessible in the real setting. Similar to the issues presented regarding training, individuals who provide supervision are not inherently present in the intervention environment (e.g., research team members). This is not realistic in most cases, and thus should receive a lower score than instances when personnel naturally present in the environment provide the supervision, or in cases where no supervision is required.

4. *Independent Identification*

This criterion captures the extent the research team was responsible for identifying individuals who would be appropriate for the interventions (e.g., through screening). As a study

moves toward effectiveness, the research team should become less involved in the selection and recruitment process. In targeted studies, having individuals naturally present in the PSUs identify and screen potential participants is more externally valid and results in a higher score on this item. Studies often are unclear, however, about whose responsibility it is to identify potentially eligible participants. For studies in which all members of the PSUs are eligible to participate (i.e., no eligibility criteria beyond provision of consent/assent), a score of *5* should be given.

5. *Independent Enrollment*

This item captures the degree to which it was the research team's responsibility to recruit and enroll individuals eligible to participate in the intervention. Once an individual completes the screen and is eligible to participate, then who is responsible for enrolling the person into the study? This information often is unclear or missing. If the intervention would be mandatory or the default activity in real life (e.g., a health curriculum integrated into regular school hours), then a score of *5* should be given. If participation in the program would be voluntary (e.g., an after-school activity, or an optional service provided by a clinic), then coders should rate the extent to which the research team enrolled participants into the intervention phase. Occasionally interventions occur in settings different from the intended environment (e.g., delivered after-school, but was intended for delivery during school). Therefore, researchers should be clear about the setting for which the program was designed, so that its effects can be correctly evaluated in relation to its level of ecological validity. For studies in which the research team is responsible for all aspects of screening and enrollment, a score of 1 should be given. However, if individuals outside of the research team (e.g., nurses, school staff, community therapists, etc.) are involved in any part of enrollment, this is considered more ecologically valid and scored based on their level of involvement.

6. *Intervention Coordination*

This item measures the extent to which the research team was involved in setting up and scheduling the intervention sessions. As programs move toward greater effectiveness and implementation in more realistic settings (e.g., schools, clinics), the research team should be less involved in all aspects, including the coordination of the intervention sessions themselves. As intervention delivery moves toward the real-world, the responsibility for coordinating providers, space, participants, etc., should be on individuals inherently present in the intervention setting. For programs incorporated directly into an existing infrastructure (e.g., administered during normal health classes by teachers or counselors), there likely is minimal coordination necessary outside of what is naturally occurring, and a code of *5* should be given.

7. *Cost of Delivery*

This criterion captures the extent to which the research team was responsible for covering expenses associated with intervention delivery. Here, it is important to consider costs that would occur if the program were delivered in the real world. Costs often associated with implementation include paying intervention facilitators, transportation, materials for participants and providers (e.g., manuals), space, snacks, etc. Although the research team often covers these kinds of expenses during efficacy trials, it is important to know if these expenses would be manageable and sustainable in the real-world setting. Additionally, in cases where the research team is not paying for these expenses, authors should report this clearly in the article. For this item, the fewer expenses covered by the research team, the higher the score given.

8. *Assessment Intensity*

This item measures the intensity of the intervention assessments. Although assessments are necessary to gather data on the effectiveness of the program, they often are incredibly time intensive and typically would not be included under realistic delivery conditions. Additionally, researchers often compensate participants for their completion of assessments during intervention studies, which introduces a risk of bias. The intensity of the assessments contributes to the program's level of ecological validity, with higher scores being associated with low intensity and less burdensome assessments. High scores on this item would be given to programs that collect study outcomes naturalistically, such as from publically available data sources that do not require consent.

9. *Inclusion/Exclusion Criteria*

This criterion evaluates the extent to which the study excluded individuals who would likely receive the intervention if broadly disseminated. This item captures the degree to which there were exclusion criteria in place beyond those needed to establish the population of interest. Coders need to distinguish between eligibility criteria designed to identify the population of interest, and criteria designed for pragmatic purposes or to maximize internal validity. A lower score should not be given when participants are excluded because they are not members of the targeted population (e.g., non-Hispanic individuals excluded from a program designed specifically for the Hispanic population). A lower score should be given, however, when participants are excluded for practical reasons (e.g., excluded 7th graders from a middle school program because their class schedule was less flexible than the other grades), or when the exclusion criteria would not likely be enforced in real-world implementation (e.g., excluded

participants from a classroom-based program because of a family history of schizophrenia). For this item, the coder should estimate the percentage of the targeted population that was not eligible due to the inclusion/exclusion criteria.

10. *Extrinsic Incentives*

This item captures the extent to which participants received extrinsic incentives that would not be available under real-world circumstances in order to encourage intervention participation. This does not include incentives given for assessment completion, which is common in prevention trials. Because research assessments likely would not be in place when interventions go to scale, providing incentives for completing assessments should not result in a lower score on this item. A score of *5* should be given if the intervention is integrated into an existing and mandatory activity (e.g., health class), as incentives over-and-above what normally would be provided (e.g., grades for the course) are most likely not needed (unless otherwise noted). If incentives are an active ingredient of the program and would be included in real-world delivery, it should be explicitly noted and taken into account when scoring this item (e.g., sticker chart).

11. *Restrictions on Outside Services during Trial*

This item measures the extent to which there were restrictions on receiving outside mental health services during the trial. Some studies may put restrictions on the services that participants can access while involved in the intervention. This can help to control for the effect of the intervention, but is not realistic for programs disseminated more broadly. Coders should rate the level of restriction that a study places on services from high (i.e., no services allowed) to low (i.e., no restrictions).

Table 1. Real-World Applicability Scale	
External Validity Criteria	
Definition of Population: Based on the inclusion/exclusion criteria and study description, what is the likely targeted population of interest?	▪ **What is the primary sampling unit (PSU; e.g., school, neighborhood, clinic, etc.)?** ▪ **What is the targeted age group?** ▪ **Is the study targeting individuals with specific risk factors?** ▪ **Is the study targeting individuals from specific demographics?**
1. *Intervention Setting:* To what extent is there access to targeted population in the intervention setting?	**1 – Highly limited access** ▪ Intervention setting has access to minimal proportion of the population of interest **2 – Limited access/access to only a very specific subgroup** ▪ Universal intervention designed for all teenagers conducted in an organization that serves a very small proportion of teenagers (e.g., boy/girl scouts) **3 – Fair access** ▪ private schools ▪ primary care clinics with restricted scope of clientele (e.g., care limited primarily to certain demographics, individuals with expensive insurance plans) **4 – Good access (most of tar. population)** ▪ single setting with broad access (e.g., primary care accepting broad range of clients, public schools) **5 – Excellent access (vast majority of tar. population)** ▪ multiple settings with excellent access (e.g., public schools and pediatrician offices)
2. *Breadth of PSU Sampling:* How broad was the initial sample from which eligible PSUs were recruited?	**1 – a single PSU in a confined area or region** ▪ a single school, neighborhood, clinic, organizational chapter **2 – multiple PSUs from a confined area or region** ▪ multiple schools/clinics within a confined area or geographic region **3 – multiple PSUs recruited from multiple areas or regions** ▪ multiple schools/clinics/neighborhoods across multiple geographic regions **4 – random sampling or probability-based weighting of multiple PSUs** ▪ simple random sampling of PSUs from population of PSUs

	▪ propensity weighting of PSUs based on the degree to which they are representative of the population **5 – inclusion of all PSUs from population of interest**
3. *Representativeness of those Screened:* Approximate Percentage of Participants within PSUs assessed for Study Inclusion Criteria? *For studies that do not require a screening process (i.e., all participants in the PSUs are eligible), code 5.*	**1 – < 20% of potentially eligible participants complete screening** **2 – approximately 20-40% of potentially eligible participants complete screening** **3 – approximately 41-60% of potentially eligible participants complete screening** **4 – approximately 61-80% of potentially eligible participants complete screening** **5 – approximately 81-100% of potentially eligible participants complete screening**
4. *Enrollment Rate for Participants Meeting Inclusion Criteria & Invited to Participate:* What percentage of participants invited to participate in the study actually enrolled in the study?	**1 – 0-20%** **2 – 21-40%** **3 – 41-60%** **4 – 61-80%** **5 – 81-100%**
5. *Evaluation of Self-Selection Bias:* To what extent did the randomized participants differ from those who were invited to participate but did not (decliners)? *If there are no decliners or 5 or the percentage of decliners is less than 5%, then code 5.*	**1 – substantial differences** ▪ differences on depression outcome variable as well as key predictors or hypothesized mediators **2 – considerable differences** ▪ differences only on depression outcome or a hypothesized mediator **3 – some differences** ▪ differences on several predictor variables or demographics characteristics (no differences on mediators or depression outcomes) **4 – minor differences** ▪ differences on 1-2 variables (e.g., demographic characteristics) that are not clearly important predictors of the outcome (no differences on hypothesized mediators or depression outcomes) **5 – no evidence of imbalance on pre-intervention characteristics**

Ecological Validity Criteria	
1. *Provider Accessibility:* To what extent were the intervention providers accessible in the intervention setting?	**1 – Minimal accessibility** ▪ Program developers **2 – Low accessibility** ▪ Academic professionals and students not inherently present in setting of interest and likely difficult to access in real life (e.g., doctoral level mental health workers, graduate trainees) **3 – Moderate accessibility** ▪ Professional community interventionists (e.g., community psychologists, social workers) **4 – High accessibility** ▪ Providers inherently present in intervention setting and readily accessible (e.g., school teachers/counselors within schools, peer interventions) **5 – No providers needed** ▪ web-based or self-paced interventions (e.g., bibliotherapy)
2. *Accessibility of Provider Trainers:* To what extent were the providers dependent on training from sources that are unlikely to be readily accessible?	**1 – Minimal accessibility** ▪ Trained directly by the intervention developers **2 – Low accessibility** ▪ Trained by academic mental health professionals or trainees (e.g., psychologists, psychiatrists, social workers) who are not inherently present in the intervention setting **3 – Moderate accessibility** ▪ Trained by professional community interventionists or non-professionals with training/experience in delivering intervention who are not inherently present in the intervention setting **4 – Good accessibility** ▪ Trained by community mental health professionals or non-professionals inherently present in the intervention setting (e.g., school counselors, teachers, community leaders) **5 – Trainers not needed** ▪ Training not required to implement intervention ▪ Self-training/online training

3. *Accessibility of Provider Supervision:* To what extent were the providers dependent on supervision from sources that are unlikely to be readily accessible?	**1 – Minimal accessibility** ▪ Supervised directly by the intervention developers **2 – Low accessibility** ▪ Supervised by academic professionals or trainees (e.g., psychologists, psychiatrists, social workers) who are not inherently present in the intervention setting **3 – Moderate accessibility** ▪ Supervised by professional community interventionists or non-professionals with training/experience in delivering intervention who are not inherently present in the intervention setting **4 – Good accessibility** ▪ Supervised by professional community interventionists professionals or non-professionals inherently present in the intervention setting (e.g., school counselors, teachers, community leaders) **5 – Supervision not needed** ▪ No supervision required
4. *Independent Identification:* To what extent did the research team identify participants who would be appropriate for the intervention (e.g., through screening)? *Code 5 for studies in which all members of the primary sampling units were eligible to participate (i.e., no eligibility criteria beyond provision of consent/assent)*	**1 – Entirely research team's responsibility** **2 – Primarily research team's responsibility** ▪ delivery entity provides the screening materials but research team conducts eligibility assessment (e.g., screening) **3 – Responsibility shared equally by research team & delivery entity** **4 – Minor support from research team** ▪ research team provides the screening materials but delivery entity conducts screening, determines eligibility, and enrolls participants **5 – Research team not involved in participant identification**
5. *Independent Enrollment:* To what extent was it the research team's responsibility to recruit and enroll individuals determined to be eligible to participate in the intervention? *If the intervention would be mandatory or the default activity in real life (e.g., a health curriculum integrated into regular school hours), then code 5. If participation would be voluntary (e.g., after-school activity, optional*	**1 – Entirely research team's responsibility** **2 – Primarily research team's responsibility** ▪ delivery entity provides lists of eligible individuals, but research team contacts them and invites them to participate, answers questions about intervention participation **3 – Responsibility shared equally by research team** **4 – Minor support from research team** ▪ research team provides recruitment materials but delivery entity contacts them

service provided by a clinic), then code the extent to which it was the research team's responsibility to enroll participants in the intervention phase.	and invites them to participate, answers questions about intervention participation **5 – Research team not involved in participant identification**
6. *Intervention coordination:* To what extent was the research team involved in setting up and scheduling intervention sessions? *Code 5 if the intervention was incorporated directly into an existing infrastructure (e.g., intervention given during normal health classes)*	**1 – Entirely research team** **2 – Primarily research team** ▪ research team coordinates space, forms intervention groups, arranges transportation; delivery entity makes reminder calls **3 – Responsibility shared equally by research team** **4 – Minor support from research team** ▪ delivery entity coordinates space, forms intervention groups, contacts families, arranges transportation; research team makes reminder calls **5 – Research team not involved**
7. *Costs of Delivery:* To what extent was the research team responsible for expenses associated with intervention delivery? *Consider costs of paying intervention facilitators, covering transportation costs, materials, space, etc.*	**1 – Entirely research team** **2 – Primarily research team** ▪ delivery entity covered only minor expenses (e.g., inexpensive materials) **3 – Responsibility shared equally by research team** **4 – Minor support from research team** ▪ research team covered minor costs such as snacks **5 – Research team covers no costs**
8. *Assessment Intensity:* How intensive were intervention assessments?	**1 – High intensity** ▪ in-person interview assessments (e.g., diagnostic interviews) & questionnaires **2 – Moderate intensity** ▪ in-person questionnaires (no interviews) **3 – Low intensity** ▪ use only assessments that are built into the intervention (e.g., activity or mood logs) **4 – Minimal intensity** ▪ participants give consent to access naturally occurring data that would not typically be available to researchers (e.g., naturalistic observations, Facebook entries, etc.), but no assessment involving direct contact **5 – No intensity** ▪ study outcomes collected from publicly available data sources that do not require consent

9. *Inclusion/exclusion criteria:* To what extent were individuals who would likely receive the intervention if broadly disseminated excluded from the trial? (To what extent were there exclusion criteria in place beyond those needed to establish the population of interest?)	**1 – Substantial portion of targeted population not represented (>20%)** **2 – Moderate portion of population not represented (>10%)** **3 – Small portion of targeted population not represented (5-9%)** **4 – Very small portion of participants excluded (< 5%)** **5 - No restrictions beyond those needed to establish a sample with the risk factor(s) of interest**
10. *Extrinsic Incentives:* To what extent were extrinsic incentives provided to encourage intervention participation (NOT assessment participation) that would likely not be available under real-world circumstances? *Code 5 if intervention was integrated into existing and mandatory activity (e.g., health class) as incentives over-and-above what would normally be provided (e.g., grades for the course) are most likely not needed*	**1 – Large incentives** ▪ large payment for intervention participation (e.g., > $100) **2 – Moderate incentives** **3 – Smaller incentives** ▪ extra-credit ▪ small gifts **4 – Minimal incentives** ▪ food/snacks at intervention sessions **5 – No incentives at all**
11. *Restrictions on Outside Services during Trial:* To what extent were there restrictions on outside mental health services during the trial?	**1 – No outside services allowed** **2 – Major restrictions** ▪ don't allow any professional mental health care outside of school counseling sessions **3 – Moderate restrictions** ▪ allow professional treatment but limit the allowable dosage (e.g., permit non-therapeutic dose or medication or a limited number of psychotherapy sessions) **4 – Minimal restrictions** ▪ allow professional treatment without dosage restrictions but do not allow a specific form of treatment (e.g., cognitive-behavioral therapy (CBT) in a CBT prevention trial) **5 – No restrictions at all**

CHAPTER III

RESULTS OF PRELIMINARY CODING

Our main research question and the reason for creating this scale was to test whether studies drop in the size of their effects as their delivery becomes more realistic. If programs with greater levels of real-world applicability show a decline in effect sizes, then it may be possible to identify exceptions to this rule, and discover what drives sustained effects across increasing levels of external and ecological validity.

To test the utility of the scale, two raters independently coded 20 studies from the depression prevention literature (see Table 2). Through the process of coding, it became clear that many items did not fall cleanly into a specific anchor point. For example, supervision of program providers varied tremendously across studies, and the descriptions of the supervisors and providers often was not clear (e.g., a clinical expert, an experienced researcher, a trained therapist). However, the primary goal of coding was for coders to maintain consistency in their ratings, such that they would be able to order studies similarly for their overall level of real-world applicability. There are several items that require quantitative information to score (e.g. *Enrollment Rate, Percent Screened,* etc.), and it is expected that coders have consensus on these items.

Table 2. Reliability by Study

Studies	ICC3	ICC2
Cardemil 2002 Study1	.84	.83
Cardemil 2002 Study2	.88	.88
Gillham 2006a	.99	.99
Gillham 2012	1.0	1.0
Gillham 2006b	.99	.99
Kindt 2014	.99	.99
Wijnhoven 2014	.99	.99
Quayle 2001	1.0	1.0
Roberts 2003	.98	.98
Yu 2002 Study 3	1.0	1.0
Gillham 2007	.94	.94
Stice 2006	.99	.99
Stice 2008	1.0	1.0
Rohde 2014	.94	.94
Clarke 1995	1.0	1.0
Garber 2009	.98	.98
Clarke 2001	.88	.88
Horowitz 2007	.99	.99
Clarke 1993 Study 1	1.0	1.0
Clarke 1993 Study 2	1.0	1.0

ICC=Intraclass correlation

Reliability

We used intraclass correlations (ICC) to assess inter-rater reliability. Because we were interested in consistency between coders rather than absolute values, we used the ICC(3,k), which is a two-way mixed-effects model that treats coders (k=2) as fixed and targets as random. For several of the items on the scale, we expected the coders to achieve absolute reliability, due to the nature of the item and concreteness of the scoring scale. For example, items such as *Enrollment Rate, Percent Screened,* and *Selection Bias* are quantitative values that should either match or not match between coders, so a consensus code would be appropriate. Therefore, we

also reported ICC(2,k) values, which is a two-way random effects model that reflects the extent of absolute agreement between coders (k=2). In most cases, these values were identical or varied only slightly. Table 2 shows the overall reliability by study and Table 3 gives the breakdown of reliability by item, as well as the minimum and maximum score given and the number of studies missing data for each item. Initially, on several items one coder scored the item as missing and the other assigned a score. In these cases, each coder went back and re-coded those items, which in all cases resulted in consensus. Recoding was done as part of the process of finalizing the scale, and our reliability reflects the consistency in ratings after coming to a consensus about whether the item was missing.

Table 3. Reliability by Items

Item Description	ICC3	ICC2	Min. score	Max score	# Missing
External Validity					
1. Intervention Setting	.74	.74	3.5	5	0
2. Breadth of PSU sampling	.87	.86	1	3.5	0
3. Representativeness of Screened	.97	.97	1	5	2
4. Enrollment Rate	.65	.63	1	5	3
5. Self-Selection Bias	.99	.99	2	5	16
Ecological Validity					
1. Provider Accessibility	.98	.98	1.5	4	0
2. Trainer Accessibility	.88	.87	1	2.5	7
3. Supervision Accessibility	.97	.96	1	2.5	6
4. Independent Identification	.98	.98	1	5	12
5. Independent Enrollment	n/a	n/a	1	5	15
6. Intervention Coordination	n/a	n/a	5	5	16
7. Costs of Delivery	n/a	n/a	n/a	n/a	20
8. Assessment Intensity	.90	.91	1	2.5	0
9. Inclusion/Exclusion Criteria	.83	.83	1	5	0
10. Extrinsic Incentives	n/a	n/a	2	5	14
11. Restrictions on Outside Services	1.0	1.0	5	5	0

n/a=not applicable

Overall, the reliability scores across studies and scale items were very good. Use of this scale requires a fair amount of judgment and knowledge of the field by coders. This could make achieving reliability difficult, as evidenced by the ICC3 score on the *Enrollment Rate* item, which was lower than the others (.65). This lower ICC resulted from discrepant scores on two studies on which the coders had different interpretations of how to compute this score. Because the purpose of this measure is to facilitate determining the overall real-world applicability of one program relative to others, discrepancies on a single item is not a major concern. The main objective is for coders to rate programs consistently, so as to be able to rank them similarly based on level of real-world applicability.

Many of the items showed a good amount of variability in scores across studies. A few items, however, had a very limited range of scores across studies (e.g., *Intervention Setting, Trainer Accessibility*, *Restrictions on Outside Services*). We only coded a relatively small number of studies; we expect that the range of the items will expand when coding larger samples of studies.

Missing information also presented a potential barrier to coding. For several items (e.g., *Independent, Independent Coordination, Costs of Delivery, and Extrinsic Incentives)*, there was too much missing data to compute a reliability score. Unfortunately, studies often do not provide the necessary information for scoring these items. Another aim of developing this scale was to highlight the kinds of information authors should provide when publishing the results of their studies to facilitate evaluation of the ecological validity of their clinical trial.

Overall, preliminary coding successfully demonstrated the utility of the scale for capturing variability in program characteristics related to real-world applicability. This initial round of

coding also highlighted potential limitations of the scale such as missing data, minimal variance on items, and vague or confusing wording of scale items.

CHAPTER IV

DISCUSSION

Evidence of the efficacy of interventions aimed at preventing depression has been increasing over the last decade. Although there has been some success with delivery of prevention programs in realistic settings, it is still unclear what factors drive these effects (Brunwasser, Kim, & Gillham, 2009; Stice et al., 2009). The purpose of the current study was to develop a measure for rating studies of randomized controlled trials with regard to their external and ecological validity. The longer term aim of this real-world applicability scale is to empirically test whether or not effects of interventions actually diminish when program delivery becomes more realistic, and to explore what factors contribute to this decline. Additionally, we can identify exceptions (i.e., programs with high internal validity and realistic delivery). Although we have formulated the anchor points of the scale items to reflect study characteristics specific to the depression prevention literature, we expect that it can be adapted to other interventions.

Results from the coding of twenty depression prevention studies indicated that most of the items captured variability in study characteristics dimensionally. Additionally, high reliability between coders suggests that the scale can be used to rank studies consistently in terms of overall level of real-world applicability. One thing that become especially apparent through the initial coding was the need for clarity and transparency in research articles. Use of vague and inconsistent terms across studies limits our ability to analyze these components accurately. When evaluating how realistic a program was delivered, it is necessary to identify all aspects of implementation for which the research team was responsible, as well as who would be

responsible for this aspect if the research team was removed. This includes factors related to costs, participant recruitment, participant enrollment, participant identification, etc. There is a tendency to look at participant factors, group factors, and dosage when examining why a program works or doesn't work, and more practical variables are not typically implicated. For example, the extent to which the research team is involved in covering the costs of program delivery or their involvement in all aspects of recruitment and implementation can impact a program's effectiveness. These hidden factors typically do not contribute to being categorized as an efficacy or effectiveness trial, but could be contributing to a program's sustained success in the real world.

There is a burden on researchers to include a lot of detailed information in a limited amount of space, and clarity to the extent that we are suggesting would likely not be compatible with page constraints. However, potential inclusion of a checklist or standardized handout with submission of an article could facilitate this process without putting a substantial burden on researchers. Additionally, one way to address this problem of limited journal space would be to place detailed materials into online supplements.

Limitations

Limitations of the scale in its current form provide important directions for future modifications. For many of the items on the scale, it was difficult to come up with anchor points that captured a wide range of possible scores from most to least realistic. In many cases, such as with program providers or supervision, it was very difficult to determine how to score an item, due to the wording or lack of complete information in the article. In these cases, coders must use their judgment, which can lead to inconsistencies across raters. Although inter-rater reliability

was good between coders who had been involved in the scale development. Future studies need to test the scale with coders naive to scale development.

Additionally, the amount of missing data was surprising and likely limit our ability to examine those factors as predictors of effects, as well as calculate reliability on those items. Nevertheless, these items are still important to include on the scale, as they are relevant to a program's real-world applicability. Moreover, the amount of missing data encountered supports the need for greater clarity in the write-up of program implementation.

Future Directions

Although the results of our preliminary coding were good, our next steps include further coding done by individuals not involved in the scale's development. Additionally, we are currently working on conducting a multilevel meta-analysis to examine the issue of diminishing effects, and to explore the relation of specific program factors to real-world applicability and effect sizes. Although we are investigating this question with regard to the depression prevention literature, we anticipate that this scale can be adapted and used more broadly to answer similar questions about other interventions and disorders.

REFERENCES

Biglan, A., Brennan, P. A., Foster, & S. L., Holder. (2004). *Helping adolescents at risk: Prevention of multiple problems of youth*. New York: Guilford.

Botvin, G. J., Schinke, S. P., Epstein, J. A., Diaz, T., & Botvin, E. M. (1995). Effectiveness of cultural focused and generic skills training approaches to alcohol and drug abuse prevention among minority adolescents: Two-year follow-up results. *Psychology of Addictive Behaviors, 9*, 183–194.

Brunswik, E. (1956). Historical and thematic relations of psychology to other sciences. *Scientific Monthly*, **83**, 151-161.

Brunwasser, S. M., Gillham, J. E., & Kim, E. S. (2009). A meta-analytic review of the Penn Resiliency Program's effect on depressive symptoms. *Journal of Consulting and Clinical Psychology*, *77*(6), 1042–1054. http://doi.org/10.1037/a0017671

Brunwasser, S. M., & Garber, J. (2015). Programs for the Prevention of Youth Depression: Evaluation of Efficacy, Effectiveness, and Readiness for Dissemination. *Journal of Clinical Child & Adolescent Psychology*, (May), 1–21. http://doi.org/10.1080/15374416.2015.1020541

Clarke, G. N., Hawkins, W., Murphy, M., Sheeber, L. B., Lewinsohn, P. M., & Seeley, J. R. (1995). Targeted prevention of unipolar depressive disorder in an at-risk sample of high school adolescents: A randomized trial of a group cognitive intervention. *Journal of American Academy of Child and Adolescent Psychiatry, 34*, 312–321

Chambers, D. a, Glasgow, R. E., & Stange, K. C. (2013). The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. *Implementation Science : IS*, *8*(1), 117. doi:10.1186/1748-5908-8-117

Flannery,D. J., Vazsonyi, A. T.,Liau, A. K.,Guo, S., Powell, K. E., Atha, H., Vesterdal, W., & Embry, D. (2003). Initial behavior outcomes for the Peace Builders universal school-based violence prevention program. *Developmental Psychology, 39*, 292–308.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., … Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*(3), 151–175. doi:10.1007/s11121-005-5553-y

Gartlehner, G., Hansen, R., Nissman, D., Lohr, K., & Carey, T. S. (2006). Criteria for distinguishing effectiveness from efficacy trials in systematic reviews. *Agency for Healthcare Research & Quality*, 1–28.

Gillham, J. E., Hamilton, J., Freres, D. R., Patton, K., & Gallop, R. (2006). Preventing depression among early adolescents in the primary care setting: A randomized controlled study of the Penn Resiliency Program. *Journal of Abnormal Child Psychology*, *34*(2), 203–219. http://doi.org/10.1007/s10802-005-9014-7

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*(9), 1322–1327. http://doi.org/10.2105/AJPH.89.9.1322

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, 34, doi:10.1007/s11121-015-0555-x

Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up. *Journal of Special Education, 36*(2), 69–79.

Hammond, K. R., & Brunswik, E. (1998). Ecological Validity: Then and Now. http://www.albany.edu/cpr/brunswik/notes/essay2.html

Higgins, J. P. T., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D., … Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. http://doi.org/10.1136/bmj.d5928

Kocsis, J. H., Gerber, A. J., Milrod, B., Roose, S. P., Barber, J., Thase, M. E., … Leon, A. C. (2010). A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Comprehensive Psychiatry*, *51*(3), 319–324. http://doi.org/10.1016/j.comppsych.2009.07.001

Mason, W. A., Fleming, C. B., Thompson, R. W., Haggerty, K. P., & Snyder, J. J. (2013). A framework for testing and promoting expanded dissemination of promising preventive interventions that are being implemented in community settings. *Prevention Science*, 1–10. http://doi.org/10.1007/s11121-013-0409-3

Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet*, *365*, 82–93.

Spoth, R., Rohrbach, L. A., Greenberg, M., Leaf, P., Brown, C. H., Fagan, A., … Hawkins, J. D. (2013). Addressing Core Challenges for the Next Generation of Type 2 Translation Research and Systems: The Translation Science to Population Impact (TSci Impact) Framework. *Prevention Science*, *14*, 319–351. doi:10.1007/s11121-012-0362-6

Stice, E., Shaw, H., Bohon, C., Marti, C. N., & Rohde, P. (2009). A meta-analytic review of depression prevention programs for children and adolescents: factors that predict magnitude of intervention effects. *Journal of Consulting and Clinical Psychology*, *77*(3), 486–503. http://doi.org/10.1037/a0015168

Thorpe, K. E., Zwarenstein, M., Oxman, A. D., Treweek, S., Furberg, C. D., Altman, D. G., … Chalkidou, K. (2009). A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Journal of Clinical Epidemiology*, *62*(5), 464–475. doi:10.1016/j.jclinepi.2008.12.011

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology*, *63*(5), 688–701. http://doi.org/10.1037/0022-006X.63.5.688

Weisz, J. R., & Jensen, a L. (2001). Child and adolescent psychotherapy in research and practice contexts: review of the evidence and suggestions for improving the field. *European Child & Adolescent Psychiatry*, *10 Suppl 1*, I12–I18. doi:10.1007/s007870170003

Weisz, J. R., Chu, B. C., & Polo, A. J. (2004). Treatment dissemination and evidence-based practice: Strengthening intervention through clinician-researcher collaboration. *Clinical Psychology: Science and Practice*, *11*, 300–307. doi:10.1093/clipsy/bph085

Winter, A. C., & Colditz, G. A. (2014). Clinical trial design in the era of comparative effectiveness research. *Open Access Journal of Clinical Trials*, *4*(6), 101–110.