2022-06

"What works" registries of interventions to improve child and youth psychosocial outcomes: a critical appraisal

Axford, Nick

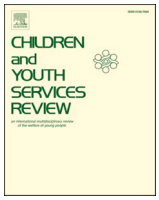
10.1016/j.childyouth.2022.106469
Children and Youth Services Review
Elsevier

Children and Youth Services Review 137 (2022) 106469

Contents lists available at ScienceDirect

Children and Youth Services Review

journal homepage: www.elsevier.com/locate/childyouth



"What works" registries of interventions to improve child and youth psychosocial outcomes: A critical appraisal

Nick Axford [a,*], Louise Morpeth [b], Gretchen Bjornstad [c], Tim Hobbs [d], Vashti Berry [c]

[a] NIHR ARC South West Peninsula (PenARC), University of Plymouth, Plymouth, UK
[b] Independent Researcher, UK
[c] NIHR ARC South West Peninsula (PenARC), University of Exeter, Exeter, UK
[d] Dartington Service Design Lab, Dartington, UK

ARTICLE INFO

Keywords:
Prevention
Early intervention
Standards of evidence
Evidence-based intervention
Registry
Clearinghouse
Implementation
Scale

ABSTRACT

The last decade or more has seen a proliferation of online registries of evidence-based interventions designed to improve child and youth psychosocial outcomes. The purpose of these resources is typically to help decision-makers make sense of the evidence and thereby inform their decision-making about investment in interventions. Most registries are underpinned by standards of evidence, which are used to guide the rating of programs by a panel of experts. While supporters extol the influence of these initiatives in terms of making commissioners more discriminating about what they invest in, detractors contend that they stifle innovation and embody an unduly narrow view of evidence and intervention. Drawing on the literature, original analysis and first-hand experience of developing, applying and using standards of evidence and associated registries, this article reflects critically on their strengths and limitations, considering issues such as focus, functionality, content, consistency and impact. It also makes proposals for developing and extending the approach, focusing on its intrinsic conceptualization of intervention development, evaluation practice and pathways to impact.

1. Introduction

The connection between evidence and policy and practice continues to be elusive. A manifestation of this is the struggle to scale evidence-based interventions (EBIs) designed to improve child and youth psychosocial outcomes. In North America they represent a small fraction of programs delivered in social care, education, health, and juvenile justice, reaching a tiny proportion of individuals and communities that could benefit (Fagan et al., 2019). The same is arguably true in Europe, where preventive interventions for health-compromising behaviors can be implemented without preliminary authorization (Faggiano et al., 2014), and Australasia, notwithstanding recent moves towards a social investment approach (SUPERU, 2016; Teager, Fox, & Stafford, 2019).

There are various barriers to identifying suitable EBIs for implementation. Study findings are not always communicated accessibly by researchers (Gorard, Griffin, & See, 2019). The number of studies can be overwhelming and challenging to understand (Gough, 2021), with decision-makers invariably lacking the time and ability to effectively identify them and interpret results (Means, Magura, Burkhardt, Schröter, & Coryn, 2015). The over-marketing of interventions with minimal or no research to support claims of effectiveness is matched by poor marketing of EBIs (Kreuter & Bernhardt, 2009). Until recently, few prevention scientists were trained in implementation science, program improvement or marketing and communication (Fagan et al., 2019). Further, decision-makers often rely on personal relationships, convenience and ideology (Gorard et al., 2019).

The last decade or more has seen increasing policy pressure to invest in evidence-based practices, especially in high-income countries experiencing public sector budget constraints. However, there is a growing awareness that while the scale-up of EBIs requires top-down statutory and philanthropic support, edicts and funding, it is necessary to have strong support from the public and system administrators and staff to produce population-wide improvements in psychosocial outcomes (Fagan et al., 2019). In other words, the challenge of scaling EBIs is about *demand* as well as supply. Increasing demand requires better communicating information about EBIs to the public and people working in public and non-governmental sectors and facilitating access to them at local, national and international levels.

The emergence of online "what works" registries, or clearinghouses, to help decision-makers understand the evidence and potentially

* Corresponding author at: University of Plymouth, N10, ITTC Building, Plymouth Science Park, Plymouth PL6 8BX, UK.
E-mail address: nick.axford@plymouth.ac.uk (N. Axford).

https://doi.org/10.1016/j.childyouth.2022.106469
Received 22 August 2020; Received in revised form 8 January 2022; Accepted 12 March 2022
Available online 16 March 2022

identify and select suitable interventions with relative ease is a response to these challenges and imperatives (Faggiano et al., 2014; Burkhardt, Schröter, Magura, Means, & Coryn, 2015; Fagan & Buchanan, 2016; Zack, Karre, Olson, & Perkins, 2019). Registries are one of several means by which knowledge intermediary organizations assemble and communicate accessible summaries of evidence to decision-makers (Gough, 2021). The term 'registry' is used to mean different things. Here, it refers specifically to online, curated and usually searchable lists of interventions that contain information about those interventions and evidence for them. Table 1 provides an overview of 24 established and active registries,[1] restricting inclusion to those published in English that include a substantial *number* of interventions focusing on children/youth and psychosocial outcomes (they may additionally include other interventions and content). Online portals that (i) focus exclusively on intervention *types* (broad approaches or modalities) and therefore tend to comprise (systematic) reviews or meta-analyses,[2] or (ii) serve only as compendiums of disparate evidence resources (e.g., guides, policy briefs, useful weblinks),[3] are out of scope in this article.

Registries are mostly underpinned by standards of evidence – criteria against which an intervention's potential usefulness is assessed. Common issues covered are intervention specificity, evaluation quality (method selection and execution), impact, and, to lesser degrees, dissemination readiness and cost (Table 2). Methods for assessing interventions against standards usually generate a global rating to aid users with minimal research methods knowledge (Maranda et al., 2021). Ratings are invariably presented in a tiered format ($\geq$2 categories), although the range varies considerably.

The rapid proliferation of registries and associated standards has created a complex and messy evidence landscape. Others have helpfully described and compared them, covering *inter alia* their aims, audiences, funding sources, search strategies, functionality and dissemination methods (Burkhardt et al., 2015; Means et al., 2015; SUPERU, 2016; Gough & White, 2018; Puttick, 2018). This article seeks to offer a fair but critical appraisal of their application in prevention and early intervention. The observations draw on the literature, our own comparison of registries, and first-hand experiences of developing, applying and using standards and registries.[4]

2. Focus

The majority of registries focus on *programs* (discrete, often manualized, packages of activity), although some additionally include broad intervention types or modalities (e.g., motivational interviewing or cognitive behavioral therapy). Focusing on programs has the advantage of pointing service commissioners and practitioners to tangible proven products that they can deliver. Paradoxically, a modality may be effective but a given individual program may not be, so care is needed in applying such findings (Means et al., 2015; Karre et al., 2017). For example, behavioral and cognitive-behavioral group-based parenting interventions are effective at improving child conduct problems (Furlong et al., 2012) but some such programs will be ineffective, or only effective in certain conditions.

There are drawbacks of focusing on programs, however. First, they are only one of multiple forms of intervention. Others include national and local policies (e.g., maximum class sizes, free school meals, measures to reduce underage alcohol consumption), whole system or place-based reforms (e.g., the Icelandic substance use prevention model), discrete units of behavioral influence or "kernels", and pathways connecting people to services (e.g., social prescribing) (Embry & Biglan, 2008; Sigfusdottir, Kristjansson, Gudmundsdottir, & Allegrante, 2011; James, James, Cowdrey, Soler, & Choke, 2013; Burton et al., 2017; Husk et al., 2020).

Second, it is increasingly recognized that complex multi-causal public health problems cannot be solved by single interventions that are often low in reach and impact; instead, a more system-orientated approach is needed (Rutter et al., 2017). This involves seeking to reshape interacting factors within the system to generate better outcomes. In turn, this demands a wider set of evaluation approaches. Existing evidence focuses largely on the effectiveness of individual interventions (typically programs) using methods grounded in linear cause-and-effect models (notably randomized controlled trials (RCTs)). This accounts, arguably, for the focus of most registries. A refocusing towards systems, however, requires a wider set of evaluation approaches, such as natural experiments, interrupted time series and simulation models (Rutter et al., 2017).

Third, evidence-based programs have poor penetration in community-based services for children and families, and there are reasons to think that this is unlikely to change significantly. The many multi-layered barriers to implementation have been documented extensively, but boil down to programs not being "system ready" and systems not being "program ready". Both issues can be addressed to some degree, but it is misguided to limit attempts to improve service quality and effectiveness to encouraging the uptake of isolated programs for specific populations (Ghate, 2016). Rather, whole system improvement is needed.

While there is a case, then, for broadening the scope of registries (and indeed some do this already[5]), this is not without challenges. Approaches that rely on meta-analyses risk mixing together disparate approaches to create an average effect size, masking the often very mixed results. Moreover, it is arguably harder to implement some of the alternative forms of intervention, notably policies or system reforms, at scale. Even the option of coding effective interventions to capture constituent behavior change techniques (e.g., Michie et al., 2013) or common elements (e.g., Leijten et al., 2019) can be problematic because individual units may not be active in isolation: the whole is more than the sum of parts. Finally, non-programmatic interventions are less amenable to controlled experimentation, with the result that there is less evidence that fits a traditional hierarchy of evidence. The use of alternative paradigms for evaluating interventions, such as an integrative theory-driven (Chen, 2015) and system (Egan et al., 2019) approaches would open the door to a wider range of intervention types, although consensus on the best methods is elusive.

3. Presentation and functionality

Registries typically present significant amounts of complex information in much simplified and accessible formats. Where available, search and comparison functions, and links to sources of useful

[1] A minority of registries are derivative, insofar as they rely on other registries for intervention reviews. They are in scope here because (i) they likely appear distinct to most registry users and (ii) they may have features relevant to this article that differ from those in the original registries (e.g., presentation and functionality).

[2] For example: National Institute for Health and Care Excellence (NICE), Campbell Collaboration, Cochrane Collaboration, Best Evidence Encyclopaedia, European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) Practice Portal, Education Endowment (EEF) Teaching and Learning Toolkit.

[3] For example: EU-Compass for Action on Mental Health and Well-being, European Crime Prevention Network, SAMHSA Evidence-based Practices Resource Center.

[4] This experience includes work with: Blueprints for Healthy Youth Development; Early Intervention Foundation 'Guidebook'; European Monitoring Centre for Drugs and Drug Addiction 'XChange' database; Evidence2Success; National Academy for Parenting Research 'Toolkit & Parenting Program Evaluation Tool'; and Washington State Institute for Public Policy Research.

[5] For example, CrimeSolutions bifurcates its toolkit into one section on programs and one on practices, and What Works for Children's Social Care (WWCSC) mixes assessments of discrete programs with assessments of broader practices or approaches.

Table 1
Overview of registries.

Registry[1]	Location	Target population	Subject focus	Programs/ modalities	Search function	Derivative?
Blueprints for Healthy Youth Development	US	Children/ youth	Prevention of antisocial behavior; promotion of healthy youth development	Programs	Yes – by outcome, target population, program type, setting, continuum of intervention, and risk and protective factors	No
California Evidence-Based Clearinghouse for Child Welfare	US	Children/ youth	Child welfare system (inc. domestic violence, substance use, behavior management, mental health)	Programs	Yes – by topic area, scientific rating, child welfare system relevance level, child welfare outcomes, age of child, program delivery options	No
Canadian Best Practices Portal	Canada	Children/ youth Adults	Public health	Both	Yes – by rating, age, gender, intervention focus, setting, determinants of health, and health promotion strategy	No
Clearinghouse for Military Family Readiness: Continuum of Evidence	US	Children/ youth (and their families)	Well-being of US military families (inc. parenting, relationships, school readiness, behavioral problems, mental health)	Programs	Yes – by 'placement' (rating), topic, target population, sector, military use, facilitator training, implementation, availability	No
Communities that Care	Australia	Children/ youth	Improving youth outcomes, preventing problem behaviors (inc. violence, harmful substance use, low academic achievement, early school leaving, sexual risk-taking)	Programs	Yes – by target age, risk factor, protective factor	Unclear
Communities for Children	Australia	Children/ youth (and their families)	Child, family and community welfare sector	Programs	Yes – by program objective, target group, keyword	No
CrimeSolutions (Youth.gov Program Directory)	US	Children/ youth Adults	Criminal justice, juvenile justice, crime victims	Both	Yes – by evidence rating, extent of evidence, topic, program type, setting, geography, age, race/ ethnicity, gender, targeted population, RCT	No
Early Intervention Foundation Guidebook	UK	Children/ youth	Early intervention (inc. mental health, maltreatment, substance abuse, crime, violence, anti-social behavior, teen pregnancy, obesity)	Programs	Yes – by evidence rating, child outcome, age group, cost rating, prevention level, setting, delivery model	No
Evidence for Impact (E4I)	UK	Children/ youth	Education (maths, reading, science, writing social-emotional)	Programs	Yes – by impact on outcomes (maths, science, social/emotional, reading, writing) at primary and secondary school levels	Unclear[2]
EPISCenter	US	Children/ youth	Youth problems such as violence, delinquency, substance use, school failure	Programs	None (list only), although matrices of (i) program × proven outcomes and (ii) program × risk and protective factors addressed	Yes, although not explicit how
European Platform for Investing in Children	Europe (EU member states)	Children/ youth	Transitions to adulthood, family-friendly workplaces, helping vulnerable children, supporting parenting	Programs	Yes – by policy category, country, evidence of effectiveness, scope of practice, type of organization implementing practice, delivery dosage, practice materials, cost information availability, evidence level	No
Evidence-based Teen Pregnancy Prevention Programs	US	Children/ youth	Teen pregnancy, sexually transmitted infections, associated sexual risk behaviors	Programs	Yes – by outcome affected, program type, target population, age group, implementation setting(s)	No
Home Visiting Evidence of Effectiveness (HomVEE)	US	Children/ youth Pregnant women	Early childhood home visiting	Programs	Yes – by meets Department of Health and Human Services criteria for an 'evidence-based early childhood home visiting service delivery model'], favorable impacts found, population served	No
National Gang Center Strategic Planning Tool	US	Children/ youth Adults (up to c.35y)	Gang problems/gang-related behavior, especially among juvenile at-risk and young adult populations	Both	Yes – by title or effectiveness code, and can order results by age range	No
OJJDP Model Programs Guide	US	Children/ youth	Juvenile justice, delinquency prevention, child protection and safety	Both	Yes – by topic, age, protective factors, risk factors, evidence rating	Yes – uses CrimeSolutions program review process, scoring instrument and evidence ratings
Pathways to Work Evidence Clearinghouse	US	Young adults/ adults (inc. parents)	Employment-focused outcomes for individuals with low incomes: increase earnings, employment, or	Programs	Yes – by target outcome(s), client characteristic, service type, state/ region where implemented, urban/	No

(*continued on next page*)

Table 1 (*continued*)

Registry[1]	Location	Target population	Subject focus	Programs/ modalities	Search function	Derivative?
			education/training; or decrease benefit receipt		rural setting where implemented, year(s) when implemented	
Pew Results First Clearinghouse	US	Children/ youth Adults	Social policy (e.g., behavioral health, criminal justice, education, public health)	Programs	Yes – by subject category, settings, rating, clearinghouse	Yes – information from 9 US clearinghouses (e.g., Blueprints, CEBC, WWC)
PracticeWise Blue Menu[3]	US	Children/ youth	Child and adolescent mental health	Both	Not available	No
Prevention Services Clearinghouse	US	Children/ youth (and their families)	Enhanced support for children and families and preventing foster care placements (inc. mental health, substance abuse prevention and treatment, in-home parent skill-based programs, kinship navigator programs)	Both	Yes – by rating and service area	No
Social Programs that Work	US	Children/ youth Adults	All areas of social policy (inc. early childhood, crime/violence prevention, education, substance abuse prevention)	Programs	Yes – by policy area	No
What Works Clearinghouse	US	Children/ youth	Education (e.g., math, literacy, behavior, early childhood)	Both	Yes – by topic (e.g., early childhood, literacy, mathematics, behavior). Can also search for reviews of individual studies	No
WSIPP Inventory	US	Children/ youth Adults	Public policy (inc. child welfare, mental health, juvenile justice, substance use)	Both	Yes – by research area (e.g., juvenile justice, child welfare, substance use disorders)	No
WWCSC Evidence Store	UK	Children/ youth	Social care (inc. domestic abuse, physical/sexual/emotional abuse, parental drug and alcohol, child mental health)	Both	Yes – by needs, service areas, effectiveness, location of evidence (UK), cost effectiveness	Yes – summaries of published systematic reviews
Xchange	Europe	Children/ youth	Substance use, youth offending and bullying	Programs	Yes – by age group, setting, outcomes targeted, risk factor	No

CEBC - California Evidence-Based Clearinghouse; OJDDP – Office of Juvenile Justice and Delinquency Prevention; WSIPP – Washington State Institute for Public Policy; WWC - What Works Clearinghouse; WWCSC – What Works for Children's Social Care

[1] See Table S1 in the Appendix for a list of registry websites.

[2] This seems to be mostly a combination of Best Evidence Encyclopaedia and the Education Endowment Foundation Toolkit.

[3] There may be more information for some criteria in the subscription-based database that informs the menu.

information (e.g., wider literature, program websites), further help registry users to select interventions. This is potentially a valuable service, as it is difficult for busy policymakers and practitioners to make sense of and apply voluminous and often complex scientific evidence (Oliver & Boaz, 2019). Moreover, social policy options are notoriously difficult to communicate owing to heterogeneous impacts on different segments of the population, multiple outcomes, long timescales and large uncertainties (Brick et al., 2018).

Despite these strengths, registries have tended not to be developed by designers working with intended end users. Relatively little is known about who registry users are, why they use registries, how they navigate and apply the information and what would best serve their needs (Burkhardt et al., 2015). What is known points to the need for innovation in content and design. Users have different levels of experience and technical expertise and use registries in different ways – to select the most promising intervention for their circumstances, or establish whether a locally-delivered intervention is evidence-based, or know whether a given intervention is likely to succeed in their setting and enhance existing provision (Means et al., 2015). There is a case, then, for multiple search methods (Zack et al., 2019) and for capturing each program's potential value to a user rather than simply comparing it to absolute standards (Burkhardt et al., 2015). Anecdotally, for example, some rating systems are hard for users to understand or interpret.

Of course, there is a tension between the amount and depth of information needed to convey the evidence base accurately, the imperative to keep content succinct and accessible and the need for extra functionality (Burkhardt et al., 2015). Applying design principles and methods through the co-production of prototypes and user experience testing could help achieve the right balance and also explore related issues, such as comparison functions (how intervention A relates to intervention B) and the positioning of information (immediately visible on the registry or buried deeper).

4. Content

4.1. Internal and external validity

Much registry content rightly concerns the veracity of claims about intervention impact, or internal validity. This helps guard against inflated claims for intervention effectiveness. Not uncommon, these reflect in part the competitive environment – developers and purveyors want to sell their interventions – and a bias against publishing null effect trials. The exaggeration of positive results manifests as misleading marketing and suspect research techniques such as outcome switching, atheoretical "fishing trips" for sub-group effects and cherry picking positive results for inclusion in abstracts (Axford, Berry, Lloyd, Hobbs, & Wyatt, 2020). Weaknesses in research design and conduct can also render evidence of effectiveness less secure than it appears; registries often decline to certify interventions, or assign them lower-than-expected ratings, because methodological flaws cast doubt on claims of effectiveness (Martin et al., 2018; Mihalic & Elliott, 2015; Steeger, Buckley, Pampel, Gust, & Hill, 2021).

Registries and associated standards tend to pay less attention to external validity or the generalizability of effects. The focus on whether an intervention was effective in the time and place it was delivered takes priority over considering if and how findings can be applied in new settings or to other populations (Cartwright & Hardie, 2012). Our analysis (Table 3) shows that while the majority of registry rating systems do account for replication of effects, how this is done varies considerably. Sometimes it is explicit, as in stipulating that there must

Table 2
Standards of evidence and ratings.

Registry	Criteria covered in standards of evidence[1]	Rating – tiers/levels	Basis for rating[2]	Openness (criteria/review process)
Blueprints for Healthy Youth Development	Intervention specificity Evaluation quality Intervention impact Dissemination readiness	3 levels: Model Plus Model Promising	1 or 2 good studies	Detailed standards and brief summary of review process
California Evidence-Based Clearinghouse for Child Welfare	Scientific Rating Scale from 1 ("strongest research evidence") to 5 ("represents a concerning practice that appears to pose substantial risk to children and families") plus an NR ("not able to be rated" because not enough research evidence)	5 levels: 1 - Well-supported by research evidence 2 - Supported by research evidence 3 - Promising research evidence 4 - Evidence fails to demonstrate effect 5 - Concerning practice	1 or 2 good studies	Fairly detailed description of (a) the scientific rating scale criteria and (b) the review and rating process
Canadian Best Practices Portal	Not explicit but takes the following into account: number of implementations; impact; adaptability/transferability; and quality of evidence	3 levels: Best practices Promising practices Aboriginal "ways tried and true"	Unclear	No information
Clearinghouse for Military Family Readiness: Continuum of Evidence	Significant effects Sustained effects Study design External replication Additional criteria (representative sample, modest attrition, practical significance, outcome measures)	4(7) levels: Effective (a) RCT (b) Quasi-experimental Promising Unclear (+, 0, -) Ineffective	1 or 2 good studies	Summary of criteria for different rating levels
Communities that Care	Not stated	In/out (only lists programs meeting criteria)	N/A (no global rating)	No information in registry, very brief description in guide
Communities for Children	Impact (positive on desired outcomes, no negative) Design (RCT, QED, high-quality qualitative, or mix) Readiness to implement in Australia (inc. training manual/documentation)	In/out (only lists programs meeting criteria)	1 or 2 good studies	Very high-level summary in 'How we select programs' of criteria and process
CrimeSolutions (Youth.gov Program Directory)	Based on (a) effectiveness and (b) strength of evidence	4 levels: Effective Promising Inconclusive evidence No effects	1 or 2 good studies	Detailed description of process and criteria
Early Intervention Foundation Guidebook	Study design Impact	5 levels: 4 - Effectiveness 3 - Efficacy 2 - Preliminary evidence NL2 - Logic model NE - No effect	1 or 2 good studies	Description of criteria and brief summary of review process
Evidence for Impact (E4I)	Study design Impact	5 levels: Strong Moderate Limited No impact Not evaluated	1 or 2 good studies	Brief description of criteria and review process
EPISCenter	Not specified per se but essentially evaluation quality, effectiveness and actual/potential replicability[3]	In/out (only lists programs meeting criteria)	Unclear	Very high-level summary of what 'evidence-based' means, with links to other registries
European Platform for Investing in Children	3 sets of criteria: Effectiveness Transferability Enduring impact	3 levels: Best practice Promising practice Emergent practice	Unclear	Summary of criteria and review process
Evidence-based Teen Pregnancy Prevention Programs	Individual studies assigned a quality rating of high, moderate, or low according to risk of bias in study's impact findings	No rating per se, but graphic display for (a) impact (positive, mixed, null, negative) on each of 5 set outcomes and (b) number of studies that this is based on (5+, 2–4, 1)	N/A (no global rating)	Detailed protocol describing review process and rating criteria
Home Visiting Evidence of Effectiveness (HomVEE)	Study design/quality and evidence of statistically significant impacts	2 levels – meets HHS criteria [Department of Health and Human Services criteria for an 'evidence-based early childhood home visiting service delivery model']: Yes ('evidence-based model') No	1 or 2 good studies	Detailed description of standards and review process
National Gang Center Strategic Planning Tool	4 dimensions of effectiveness: Conceptual framework Program fidelity Evaluation design Outcome evidence	3 levels of effectiveness: Effective/exemplary Effective Promising	Other (points system)	High-level summary of criteria and scoring system

(*continued on next page*)

Table 2 (*continued*)

Registry	Criteria covered in standards of evidence[1]	Rating – tiers/levels	Basis for rating[2]	Openness (criteria/review process)
OJJDP Model Programs Guide	Based on (a) effectiveness and (b) strength of evidence	3 levels: Effective Promising No effect	1 or 2 good studies [based on CrimeSolutions]	Directs reader to CrimeSolutions, which provides detailed information about standards and process
Pathways to Work Evidence Clearinghouse	Strength of evidence (essentially study design/quality and effectiveness)	6 ratings for program effectiveness for each of 4 set outcomes: Well supported Supported Mixed support Not supported Insufficient evidence No evidence Individual study quality rated high, moderate, or low	1 or 2 good studies	Protocol offers detailed description of criteria and process
Pew Results First Clearinghouse	Quality of evidence and nature of impact	5 ratings: highest rated, second highest rated, mixed effects, no effects, negative effects, insufficient evidence (each program receives as many ratings as clearinghouses have rated it)	N/A (reports other registries' ratings)	Brief summary of how it collates information and ratings from other clearinghouses
PracticeWise Blue Menu[4]	Strength of evidence (essentially study design and impact)	5 levels: Best support Good support Moderate support Minimal support No support	1 or 2 good studies	Minimal information
Prevention Services Clearinghouse	Study design/execution and effectiveness	4 tiers (global rating for program/service): Well-supported Supported Promising Does not currently meet criteria Individual (eligible) studies rated high, moderate, or low	1 or 2 good studies (for global rating) plus systematic review (meta-analysis) for effect size	Detailed description of standards and procedures
Social Programs that Work	Study quality and effectiveness	3 levels: Top tier Near top tier Suggestive tier	1 or 2 good studies	No information (points to 'related resources', which include approaches to assessing the quality of an RCT)
What Works Clearinghouse	Effectiveness for specified outcomes, and extent of evidence	No program rating per se. 3 levels for whether study meets WWC design standards: meets without reservations; meets with reservations; does not meet. 6 levels[5] for program effectiveness by outcome domain: positive; potentially positive; mixed; no discernible effects; potentially negative; negative	No global rating but systematic review (meta-analysis) for effectiveness	Very detailed description of review process and criteria, with extensive publicly accessible guidance (documents, videos, webinars) for reviewers, study authors and registry users
WSIPP Inventory	To be included in a meta-analysis, an evaluation must either have a control or comparison group or use advanced statistical methods to control for unobserved variables or reverse causality	None – results focus on effect size from meta-analysis and benefit-cost ratio	N/A (no global rating)	Detailed description of cost-benefit model and process for applying it
WWCSC Evidence Store	EMMIE: Effect Mechanisms Moderators Implementation Economic impact	Overall effectiveness (5-point scale, from 'Negative effect' to 'Consistently positive effect') Strength of evidence (5-point scale, from 'Very low' to 'Very high')	Systematic review	High-level summary with links to other registries for (a) EMMIE criteria and (b) definition of an 'acceptable study'
Xchange	Intervention definition Evaluation quality Impact	6 levels: Beneficial Likely to be beneficial Possibly beneficial Additional studies recommended Unlikely to be beneficial Possibly harmful	1 or 2 good studies	Detailed description of review process and criteria

[1] In some instances replication potential is not stated explicitly as a criterion but instead is included within effectiveness when higher ratings require evidence of effectiveness from two or more studies (see Table 3).

[2] The term "1 or 2 good studies" is borrowed (see Gough & White, 2018; Gough, 2021) and conveys the idea that a rating can be obtained based on a small number of studies – usually 1 or 2 (occasionally more) and often a subset of all evaluations of the program (those deemed to be better quality or more rigorous).

[3] Refers to "Evidence-based" (meaning rigorously evaluated and shown to work) and states that included programs "tend" to have been "assessed in large studies with diverse populations or through multiple replications by independent researchers".

[4] There may be more information for some criteria in the subscription-based database that informs the menu.

[5] The number and labels of levels differs depending on the source but 6 is the largest and most comprehensive.

be evidence of effectiveness from two or more high-quality studies with non-overlapping samples to achieve the highest rating. Elsewhere it is more subtle, for instance using icons to show whether single or multiple studies inform a rating, or simply stating that consistency of effects across studies influence ratings. Only two registries[6] specify required characteristics of study samples (both based on geography/culture).

Other registry content relevant to external validity varies widely in amount (from none to lots) and focus. Most common is information about (i) the location (country, state, city) where a program has been evaluated or implemented, and (ii) the demographics of study samples (age, gender, ethnicity; less commonly socio-economic status (SES), education level, special educational needs (SEN) status, English learners) at a program or study level. Occasionally, registry users are advised to consider factors that will affect program suitability in a host community and fit with the organization, including whether it targets relevant risk factors.

Our findings are supported by an analysis of the extent to which registries record context-specific implementation factors that affect intervention outcomes (Horne, 2017). The study rated reports on all 55 youth development programs in the top evidence category in seven major US-based registries. Nearly all reports (91%) provided context-specific information about participants, but far fewer commented on issues such as fidelity/adaptation (55%), the wider service environment (37%), quality assurance methods (22%), organizational leadership and administrative support (19%), and the demographics, education level and turnover of staff (15%). Moreover, content was primarily descriptive, with little on causal relationships between implementation and outcomes. Registries were deemed to provide insufficient information to guide context-sensitive decision-making about program replication and adaptation. Tellingly, the study found that relevant data and findings were often present in the original evaluation write-ups.

This presents a challenge. Decision-makers need to know: will the intervention be effective in my community, implemented by my organization, offered to my clients and run by my staff? Yet positive effects identified in an initial effectiveness trial often fail to materialize in subsequent trials. A stark manifestation if this is several flagship EBIs from the US struggling to produce positive effects in Europe (e.g., Sundell et al., 2008; Skärstrand, Sundell, & Andréasson, 2013; Baldus et al., 2016; Berry et al., 2016; Humayun et al., 2017; Fonagy et al., 2018). Design issues in the original and replication studies might account partly for this phenomenon but there are also contextual and implementation-related explanations – the differential quality of services as usual, adaptations removing active ingredients, compromised fidelity, poor fit with local systems and culture, and possibly different aetiological mechanisms underlying adverse outcomes (Burkhart et al., 2019).

Notwithstanding efforts by some registries to address the issue, more could be done. Registries could report more about where and how the intervention was implemented when found to be effective, how contextual factors influenced implementation and outcomes, and for whom and through which mechanisms the intervention was effective (possibly drawing on study designs besides RCTs and quasi-experimental design (QED) studies). Of course, decision-makers must still apply this information in context and may need support to assess implementation capacity and the potential of a given intervention to address relevant mechanisms operating locally.

4.2. Implementation readiness and experience

While the majority of registries provide *some* information on intervention set-up and implementation, the nature and amount are variable (Table 3). Often it is cursory, in some cases consisting only of the program purveyor's website address or details on how to order materials. It is also primarily descriptive, notably where the program has been implemented, delivery requirements (e.g., staffing qualifications) and what support is available (e.g., manuals, training, fidelity measures). Costs are given infrequently, and when they are the information tends to be brief and not in disaggregated form. Implementation readiness is barely assessed and tends not to be part of the overall rating; only exceptionally is it an entry criterion, whether minimally (the program must be active) or based on a detailed assessment. The reporting of practitioners' implementation experiences is rare. Some registries provide generic (rather than program-specific) guides on implementation. The net result is that registry users can end up enlightened to the n^{th} degree about intervention effectiveness in a given setting but in the dark about how easy it is to deliver, what factors enable or hinder this, and whether providers and users actually like the intervention (Neuhoff, Axworthy, Glazer, & Berfond, 2015).

Empirically, the impact of this deficit is unknown, but it seems likely that some effective interventions are passed over for want of information, or selected but then discarded as the lack of delivery infrastructure becomes apparent, or adopted but encounter implementation difficulties because of unresolved issues. This wastes time and effort, leading potentially to registry users becoming disillusioned with registries and even evidence-based interventions *per se*. A dissemination readiness criterion is therefore essential; interventions lacking this should not be recommended (Fagan & Buchanan, 2016; Buckley, Fagan, Pampel, & Hill, 2020). Providers' experiences of delivering the intervention, including barriers and solutions, should also be included, ideally organized using recognized implementation science frameworks (see Nilsen, 2020).

4.3. Effectiveness

Registries tend to rate programs on the overall strength of evidence, combining methodological quality with effect on outcomes, rather than focus on the size of intervention effects. Effect sizes are sometimes, though inconsistently, given in narrative study descriptions, and occasionally a registry reports a meta-analysis of program studies (Table 3). They rarely affect the overall rating (according to reported standards, at least).

Yet without effect size information, commissioners may select – all else being equal – a less effective program over one that is more effective. This may lead to disappointment, as it is easy to assume that a high strength of evidence rating implies substantive meaningful effects when in fact they could be small (Means et al., 2015). There is a case, then, for reporting effect sizes so that registry users can compare them for different programs.

Some caution is needed, however. First, not all outcomes are of equal value: a small shift on behavior may be worth more than a larger shift in knowledge or attitudes. It is important to convey the public health benefit of the outcome in question. Second, there is a risk of devaluing apparently small effects generated by *universal* interventions, which can have more population relevance than the larger effect of a targeted intervention (Greenberg & Abenavoli, 2017; Tanner-Smith, Durlak, & Marx, 2018). Third, some methodological factors are associated with effect size. On average, effects are smaller in higher-quality trials, larger studies, trials with stronger counterfactuals and evaluations with no developer involvement (e.g., Eisner, 2009). Fourth, methods of communicating effect size affect users' engagement with the data and their perception of effectiveness (Lortie-Forgues, Na Sio, & Inglis, 2021). Care is therefore needed to convey effect sizes accurately and accessibly, so that users do not swich off, expect greater effectiveness than is realistic, or prematurely reject beneficial interventions. Multiple metrics alongside guidance may be optimal.

Most registries also adopt a somewhat reductionist approach to effectiveness. They aggregate across studies to derive an impression of effectiveness given the preponderance of evidence. There is less consideration of why trials yield different outcomes in different contexts, and therefore where and for whom a given intervention might be

[6] Communities for Children and Xchange.

Table 3
Information about intervention effectiveness and implementation.

Registry	Effect size	External validity	Implementation information
Blueprints for Healthy Youth Development	Partial – where included (does not apply to every program), either reports effect sizes from primary studies, or summarizes them (e.g., small, medium, large), or reports third party meta-analyses	States briefly where studies were conducted, and summarizes demographic information from studies (race, ethnicity, gender). Top two rating tiers require evidence of effectiveness from 2 well-conducted RCTs or 1 high-quality RCT and 1 high-quality QED evaluation	Describes program-specific training and technical assistance required/available. Programs only approved if "dissemination ready"
California Evidence-Based Clearinghouse for Child Welfare	No information	Brief information about study sample (age, gender, race/ethnicity) and location. Top rating tier requires evidence of effectiveness from ≥2 RCTs with non-overlapping samples	Gives (a) program-specific reference(s) for manual (s), availability of training, information from program representative (e.g., formal support, fidelity measures, manuals, cost), and (b) links to general implementation tools and resources
Canadian Best Practices Portal	No information	Top rating tier means that intervention has, through multiple implementations, demonstrated high impact (positive changes related to desired goals) and high adaptability (successful adaptation to different settings)	Program-specific headings for implementation history, expertise required for implementation, supports available for implementation, available resources/products
Clearinghouse for Military Family Readiness: Continuum of Evidence	No information	Section in some program write-ups titled 'Previous use'. Top rating tier ('effective') requires evidence of ≥1 successful external replication(s)	Brief program-specific information on implementation considerations, training, and cost
Communities that Care	No information	No program-specific information. General guides encourage communities to consider factors that will affect whether a program is suitable for their community	Very brief program-specific "monitoring recommendations" and "implementation tips". General guides that cover implementation-related issues. Criteria for inclusion of programs include (a) feasibility for implementation and monitoring in Australia, and (b) availability of support and advice to assist Australian implementations
Communities for Children	No information	Limited program-specific information about where programs have been evaluated and found to be effective. Guide to selecting programs advises attention to fit with organization, target audience and risk factors to be addressed. Inclusion criteria require that program has been evaluated in a cultural setting that is similar to Australia	Brief program-specific information on training requirements and cost. Implementation readiness criteria are part of standards. Accompanying guide has sections on "Resourcing requirements" and "Preparing staff to deliver the program"
CrimeSolutions (Youth.gov Program Directory)	Effect size affects rating but effect size information not presented	Brief information about study settings and study sample demographics. Icons distinguish between programs that have been evaluated with single or multiple samples, although this does not influence rating per se	Brief program-specific information about materials and training
Early Intervention Foundation Guidebook	No information	Filter allows selection of programs implemented in the UK. States countries where program has been implemented and evaluated. Brief narrative description of study sample demographics (inc. age, gender, eligibility for free/reduced school meals). Top rating tier requires evidence from ≥ 2 high-quality evaluations demonstrating positive impacts across populations and environments	Program-specific summary of implementation requirements (who can deliver it, training requirements, how practitioners are supervised, systems for maintaining fidelity) and licensing requirement
Evidence for Impact (E4I)	Presents mean effect size based on meta-analysis of relevant studies	States where program has been evaluated and whether evaluated in the UK. Top rating tier interpreted as "has been shown to work in many well-controlled studies", although technically only requires ≥ 1 RCT	Brief program-specific information on staffing requirements and professional development/ training needed
EPISCenter	No information	Single rating of "evidence-based" tends to mean that the intervention has been assessed in large studies with diverse populations or through multiple replications	Videos and materials about aspects of implementation, and multiple program-specific documents (e.g., FAQs, tools for planning/ readiness, data collection, fidelity monitoring)
European Platform for Investing in Children	Not effect size per se, but reports scores for treatment and control groups for measures and divides these into outcomes with (a) effects and (b) no effects	States age of study samples and countries where (a) implemented and (b) found to be effective. Transferability criteria (requiring evidence of a positive effect in a robust study in ≥1 additional population(s) beyond the original study population) influence rating	Very brief program-specific information on practice materials and cost, with link(s) to available resources
Evidence-based Teen Pregnancy Prevention Programs	Provided inconsistently (i.e., for some outcomes in some studies)	Summarizes where study was conducted and sample demographics (age, race/ethnicity, gender), also states where program has been implemented (country, state, city). Number of studies contributing evidence to a given outcome reflected in size of icon	Program-specific implementation readiness score based on assessment of (a) curriculum and materials, (b) training and staff support, and (c) fidelity monitoring tools and resources. Also program-specific information about staffing, materials/resources, additional needs for implementation, fidelity, training and staff support, and allowable adaptations. Generic section on "experiences in implementation of evidence-based programs" (interviews with developers, and success stories).

(*continued on next page*)

Table 3 (*continued*)

Registry	Effect size	External validity	Implementation information
Home Visiting Evidence of Effectiveness (HomVEE)	Calculates effect size for each outcome measured in each included study	Summarizes location, setting and sample demographics (gender, age, race/ethnicity) for the studies reviewed. Replication of effects (in multiple studies with non-overlapping analytic samples) partially informs rating	Program-specific information about implementation prerequisites, training requirements, estimated costs, adaptations/ enhancements, and implementation experiences (drawn from studies)
National Gang Center Strategic Planning Tool	No information	No information, and does not obviously affect rating	Contact details of program developer/purveyor, and limited recommendations for strategies and practices for different age groups
OJJDP Model Programs Guide	No information, although effect size influences rating	Brief description of study location and setting, also sample demographics (age, gender, race/ ethnicity). Icons distinguish between programs that have been evaluated with single or multiple samples, although this does not influence rating per se	Summary of cost and availability of program materials. Links to implementation guides on diversion, juvenile re-entry, and school-based bullying prevention (focus on pre-implementation stage)
Pathways to Work Evidence Clearinghouse	Gives percentage point change (employment, training) or $ per year change (earnings, benefit receipt), plus effect in standard deviations (all)	For each program states implementing organization(s), state/region where implemented, staffing, local context, populations served (age, gender, SES, education level, race/ ethnicity), and funding source. Top rating tier requires $\geq$ 2 impact studies of moderate or high quality showing evidence of favorable findings within the outcome domain	Description of aspects of how intervention was implemented at the time of the evaluation (see 'external validity'), also fidelity measures and cost information
Pew Results First Clearinghouse	Varies depending on clearinghouse	Varies depending on clearinghouse	Varies depending on clearinghouse
PracticeWise Blue Menu[1]	No information	No program-specific information. Top rating tier requires evidence from $\geq$2 RCTs demonstrating efficacy	No information
Prevention Services Clearinghouse	Gives effect size and implied percentile effect by outcome, both overall and by individual study	No program-specific information. Top rating tier requires evidence of favorable effects from $\geq$ 2 high/moderate-quality studies with non-overlapping samples	Basic program-specific information on: dosage; location/delivery setting; education, certification, and training; program or service documentation; and developer contact details
Social Programs that Work	Narrative description (e.g., % increase or decrease)	Brief narrative description of study sites and sample demographics. Top rating tier requires positive effects in $\geq$2 RCTs conducted in different implementation sites, or, alternatively, in 1 large multi-site RCT	Website addresses of respective programs
What Works Clearinghouse	Calculates effect size and "improvement index" score[2] for each outcome in each included study. Improvement score also provided at program level (by outcome)	Summarizes location (country, state, urbanicity), setting and sample demographics for studies meeting WWC design standards (inc. grades, race, gender, free/reduced price lunch, special education, English learners). Replicability per se does not obviously affect rating, although consistency of findings across studies is taken into account	States program-specific costs by ingredient (personnel, facilities, equipment/materials). Describes training and support provided in included studies
WSIPP Inventory	Calculates effect sizes for relevant outcomes based on meta-analysis of relevant studies	No information. Cost-benefit model is based on state of Washington	Detailed program-specific cost-benefit analyses
WWCSC Evidence Store	No information	Lists countries where program has been implemented. Ratings take account of consistency of effects across studies	Brief program-specific summary of (a) who can deliver it, (b) training and supervision requirements, and (c) what supports good implementation
Xchange	Included in narrative description of study results (only European studies described)	Lists countries where program has been evaluated, and states ages of study participants (only European studies described). Ratings take into account evidence of replication of effects (top rating tier requires positive effects in $\geq$2 studies in Europe)	Minimum entry criteria include evidence that program is active or able to be used in Europe. Provides program-specific information from providers in Europe about implementation experiences (main obstacles, how obstacles were overcome, lessons learnt, strengths, weaknesses, opportunities, threats, recommendations)

[1] There may be more information for some criteria in the subscription-based database that informs the menu.
[2] Expected change in percentile rank for an average comparison group student if that student had received the intervention.

suitable. There have been moves towards a more "realist" approach in methods of developing and evaluating complex interventions, including trials (Fletcher et al., 2016). As more such studies emerge, registries can present a more nuanced picture of effectiveness.[7] In the meantime, registries could make it easier for users to search by the contexts and populations in and with which interventions have demonstrated effectiveness.

Registries also vary in whether they report interventions found to have null or negative effects, with many opting not to do so. This might reflect a reticence to discourage intervention development and evaluation or to deal with disgruntled (and possibly litigious) developers and purveyors. It may also be hard to identify relevant programs because publication bias favors positive effect trials. However, publishing such information could help to eliminate ineffective and iatrogenic interventions (Fagan & Buchanan, 2016).

[7] Registries that use the EMMIE (Effect, Mechanisms, Moderators, Implementation, Economic impact; Johnson et al., 2015) standards do this to some extent already (e.g., WWCSC Evidence Store).

4.4. Dynamism

Despite best efforts, which include updating reviews as new studies become available or at set junctures, registries struggle to stay current. There are several reasons for this. First, the design and packaging of interventions changes. Developers sometimes add content or adjust the duration, perhaps to reflect emerging evidence from other studies or implementation experiences. They may also subtract or adapt elements to facilitate scale-up, for instance replacing in-person training with virtual/digital options. Second, it can be difficult to keep abreast of the developing evidence for interventions. New studies take time to identify and process, which is problematic if they would change the rating; replication studies tend to yield more equivocal results. Third, the wider evidence base may advance but interventions that fail to evolve accordingly can quickly become outdated. Fourth, when services as usual improve, an intervention that was effective historically may no longer produce significant added value – the so-called "rising tide" phenomenon (Chen, 2015). Fifth, standards of evidence are rightly updated as the evidence-based practice movement progresses and high-quality evaluations become more commonplace (Fagan & Buchanan, 2016), but it means that some studies that secured a high rating in the past would do so no longer.

Collectively, these challenges can cause registry ratings to become detached from interventions in their current incarnation – what was assessed no longer exists – or anachronistic in the new context. While this is clearly misleading, it is a difficult problem for registries to address with limited resources. For instance, it would be a mammoth undertaking to re-review earlier studies against updated standards. What should be straightforward is to state clearly when ratings were made and using what version of standards, to explain when and why standards are updated, and to provide guidance on interpreting and acting on ratings. More ambitiously, the better use of resources across registries (see Section 7 below) would assist with keeping registry content current.

5. Rating systems

5.1. Openness

In most registries, each intervention receives a global rating. The exact rating system and terminology vary considerably, with narrative, numerical and semiotic approaches in use, and approaches can appear complex to the uninitiated if they combine different dimensions of criteria or seek to capture nuance. In theory, however, the use of a clear and consistent yardstick permits the comparison of options within a registry, which in turn can influence decisions about what to implement, avoid or decommission.

While this is helpful, there is a case for greater openness about the processes and reasoning that lead to intervention ratings. An independent reviewer should arrive at the same rating as the registry based on what they know about registry protocols (Burkhardt et al., 2015). The criteria used to inform ratings are usually published on registry websites but this is insufficient to replicate assessments externally (especially where they are in summary form only). Detailed protocols giving step-by-step instructions in applying standards are less common, as are public-facing justifications for ratings (Table 2). While attractively simple, ratings mask complex decision-making that might be of interest and use for some registry users. Besides stated criteria, ratings are affected by organizational contexts, the comprehensiveness of the search for relevant studies and the extent to which reviewer judgment is permitted (Burkhardt et al., 2015). Ratings are also affected by expert panel members' expertise, ideology and perspectives and less tangible factors such as their preparedness for meetings. In addition to documenting criteria and review processes, then, registries could usefully state which criteria were fulfilled to meet the current classification, the date of review and the steps needed to gain a higher rating (Karre et al., 2017).

Counterarguments to providing such details are that they are of little interest beyond developers, purveyors and the research community, would overwhelm most registry users, and are unlikely to drive better decision-making. Moreover, the process of reviewing and rating programs cannot be completely objective: some inconsistency is inevitable. On balance, however, greater openness would arguably enhance the perceived trustworthiness of registries, in turn boosting their use.

5.2. Thresholds

Registries vary in the criteria they use and how those criteria are applied, including the thresholds between one rating category and those above or below. Concerns have been raised about registries that appear to award interventions high ratings based on one or two "good studies" – in terms of rigour or quality – showing positive effects while ignoring a potentially wider evidence base (Gough & White, 2018; Gough, 2021). Our analysis of registries suggests that rating criteria referring to evidence from a few selected studies is the dominant approach (Table 2). It is less clear, however, that registries deliberately exclude from consideration other well-conducted studies that did not find a positive effect. As such, the focus on one or two studies to inform ratings may be a problem more of how standards are expressed than how they are applied. That is, registries often identify *all* studies of an intervention before filtering some out to concentrate on only the best or most relevant; panels then make judgments based on the *preponderance* of evidence, taking study quality and the number, nature, and sizes of effects into account. Robust studies tend to count for more than weaker studies, but robust studies showing positive results tend not to count for more than robust studies with null or negative results. Admittedly, this could often be clearer. The danger of overlooking robust studies showing null or negative effects might also be countered by undertaking a statistical synthesis (or meta-analysis) of all relevant studies, assuming they are sufficiently homogeneous, and weighting by sample size, although this is not possible for the many interventions evaluated only once or more than once but with different outcomes being measured.

Some registries do explicitly draw on a wider set of studies of the intervention in question to provide useful information about issues such as context, moderators and implementation, although they tend not to inform ratings (or at least not obviously). There is a strong case for systematically reviewing and synthesizing the entire known relevant knowledge base for a given intervention to avoid potential bias in ratings (Gough, 2021); if this exceeds available resources, it seems essential to at least cite the primary evidence used to determine a rating (the robust study or studies that met the registry's standards to determine a positive rating) and be explicit how these studies differ from the additional evidence base (particularly robust designs that show null or negative effects).

5.3. Proving or improving?

More inclusive tiered rating systems have significant strengths as regards intervention development and evaluation. They assign value to interventions that have not been tested in a comparison group study, thereby recognizing the value of a robust intervention design and other types of evaluation (feasibility studies, pilot trials, and pre-post designs). They also give pointers to intervention developers and evaluators about possible next steps as regards design and evaluation.

However, they can have unfortunate side-effects. One is to imply that intervention development is necessarily linear, culminating in proof of effectiveness via one or more RCTs followed by implementation and scale-up. This "pipeline paradigm" (Knox, Hill, & Berlin, 2018) is arguably the default model in prevention and early intervention for achieving impact (cf. Asmussen, Brims, & McBride, 2019). Reality is rarely so orderly, though, with interventions and their evaluations invariably evolving more organically. Another drawback is conflating movement up a registry rating scale with improvement to the

intervention. Achieving a higher rating, even demonstrable effectiveness, does not necessarily signal improvement to the intervention *per se*; there could still be major issues with, say, its ability to engage marginalized groups, or the quality of provider-user interaction. Proving is not improving. Equally, improvement is not contingent on jumping rating levels. Interventions can be strengthened on multiple fronts – theory of change, specification of core and flexible elements, effective targeting of suitable participants, monitoring of outcomes, quality of technical assistance – and the value of those improvements can be tested using various methods (Lemire, Christie, & Inkelas, 2017). Such changes and evidence of their value may not show up on a registry rating system.

5.4. No-man's land

In registry rating systems informed by the pipeline paradigm, there is often no recognition of any study "below" a good efficacy trial, or, in tiered approaches, a chasm between the lower levels, typified by simple pre-post evaluations, and the higher levels, which always require an RCT or QED. This gap is unhelpful for practice because it leaves the majority of interventions stranded in a no-man's land between initial evidence of promise and enthusiastic endorsement by registries, even if they clearly embody the features of "proven" interventions. It is also methodologically dubious; without detracting from the necessity and value of RCTs in the right conditions, it is neither possible, desirable nor necessary to test all interventions experimentally, for example because of a lack of equipoise or resource, or because the intervention is not sufficiently developed. Yet there are (underused) means of strengthening causal inference in non-experimental studies. Quantitative techniques to mimic control groups include algorithms based on epidemiological data (Ford, Hutchings, Bywater, Goodman, & Goodman, 2009) and statistically derived controls using government administrative data (Adler & Coulson, 2016; Piazza, Corry, Noble, & Bagwell, 2019).[8] There are also qualitative methods that seek to validate the intervention theory of change (or rule out competing hypothesized causal mechanisms) or explore stakeholders' perceptions of causal relationships (Stern et al., 2012; White & Phillips, 2012).

Standards of evidence arguably need some refocusing accordingly. Currently, most do not recognize qualitative evidence of program impact owing to a lack of available protocols for assessing qualitative evidence (Means et al., 2015). There is also a need routinely to give due weight to "best possible" evidence, for example by allowing non-randomized designs to achieve high ratings when a trial would be unethical or infeasible (Movsisyan, Melendez-Torres, & Montgomery, 2015). This applies especially given that certain types of QED may yield effect estimates that are consistent with those obtained in trials. Credit might also be given to interventions that have not been trialled but nevertheless resemble those that have (and been shown to be effective) in terms of content and form. There is a good case for encouraging providers to integrate common elements of effective practice into existing interventions alongside the adoption of branded programs (Lipsey, 2020). When this is done well, the case for conducting a new RCT is diluted.

6. Impact

6.1. Promoting "what works"?

Although registries plausibly contribute to the greater implementation of EBIs at the expense of interventions with no or disappointing evidence of effectiveness, there is – ironically – little robust empirical data on trends in this respect or their causes. That said, the signs are not promising. Few interventions with high ratings are scaled in North America or Europe, while many interventions with low or no ratings are prevalent. An analysis of children's centers in the UK, for instance, found that a minority offered EBIs, and of these a minority offered them in full (Goff et al., 2013). More recently, it was estimated that less than 0.1% of expenditure on children's services in Northern Ireland went to recognized EBIs (Kemp, Ohlson, Raja, Morpeth, & Axford, 2018). Indeed, there is now evidence that the scaling trajectories of a significant proportion of public health interventions bypass efficacy and/or real world effectiveness testing (Indig, Lee, Grunseit, Milat, & Bauman, 2018).

It might be countered that it is too early to detect the impact of registries, although this ignores how long some have existed (over 20 years). Another defence, namely that registries should not be judged by the uptake of EBIs because that lies beyond their mission, overlooks the explicit aspirations of many. More plausibly, registries are not being used as intended by people with the power to influence service provision. For example, a recent US study analyzed the extent to which policymakers in state statutory agencies responsible for behavioral healthcare promote the use of registries by referencing them on their websites (Maranda et al., 2021). The absolute number of references was low and three out of the 28 registries it looked at accounted for 74% of references. The study also considered factors that might affect usage, including registry features (e.g., longevity, rating system usefulness, value-added options such as planning guides) and contextual factors, such as local legal or funding requirements to select interventions from a designated list. Others have argued that there is a tendency for intermediary organizations that typically host registries to focus on promoting engagement with evidence at the expense of supporting the application of evidence in decision-making (Gough, Maidment, & Sharples, 2018).

Meanwhile, perverse effects of registries should also be countenanced. The longer lists of interventions in more inclusive registries may increase the risk of interventions with limited or no ability to improve outcomes being selected and implemented, which wastes time and resources and undermines public confidence in science if the expected results are not evidenced (Fagan & Buchanan, 2016). Certainly there is anecdotal evidence of programs being listed conferring credibility in the eyes of commissioners and of developers commending their intervention based on it being "on the list", regardless of its rating and the availability of better-evidenced alternatives. There is perhaps little that registries can do about this besides continuing to issue health warnings; it does, however, point to the need to build capacity for more informed and intelligent commissioning.

6.2. Evaluation practice

Standards of evidence used to assess interventions for registries have the potential to exert a healthy influence on evaluation conduct and the accuracy and transparency with which studies are reported. Experience suggests that investigators and developers are increasingly concerned to meet accepted quality criteria, thereby maximizing the likelihood of the intervention in question receiving a high registry rating (subject to positive results). Registries and other industry standards could be mutually reinforcing, for instance in relation to expectations to preregister studies and share research code, data and materials (e.g., Kidwell et al., 2016; Gennetian, Tamis-Lemonda, & Frank, 2020). Without empirical data, however, it is difficult to disentangle the contribution of registry standards from related influences, such as the increased expectation to report trial results in compliance with CONSORT guidelines (Moher et al., 2010). Moreover, the continued sub-optimal design and conduct of many studies reviewed has prompted some registries to produce guidance on how to avoid common methodological pitfalls that block interventions receiving a higher rating (e.g., Martin et al., 2018).

Standards of evidence and registries may also have unwelcome effects on evaluation practice. First, they can disincentivize the continued testing of an intervention that reaches the pinnacle of a given rating system. This is partly because of the cost and effort involved,

[8] Some of these may, in some registry rating systems, be accepted as a QED to the extent that they create a valid counterfactual.

particularly if there seems little obvious benefit to be gained, but more that there is potentially much for developers and purveyors to lose; a new trial with null or negative effects could result in removal of the intervention from a registry or a significant fall in rating (Karre et al., 2017), either of which could be detrimental to the likelihood of the intervention being commissioned. To mitigate this, some registries add time as a criterion, so that a program can only achieve the highest rating if it was robustly evaluated in, say, the last five years.

Second, registries may incentivize the wrong kind of evaluation. Ideally, evaluation would be used to help improve interventions, grounded in an internally-developed, multi-year and proportionate roadmap for evidence generation to optimize delivery and impact (Brooks, Boulay, & Maynard, 2019). Instead, developers and purveyors can be inclined to chase *external* endorsement, especially if a certain rating is thought to be positively associated with being commissioned. This can lead to interventions being trialled prematurely or unnecessarily, increasing the likelihood of uninformative null or negative results (Axford et al., 2020). A more formative or developmental approach (Patton, 2010) is needed, which requires refocusing standards grounded in a more summative approach.

Third, registries may incentivize not evaluating at all; since many interventions are commissioned regardless of whether they appear on a registry, it may be safer to remain "off list" rather than risk a low rating. This reinforces the case for registries listing interventions with no evidence on impact.

6.3. Appraising existing practice

Standards and registries can be used to make a realistic appraisal of the nature and quality of existing practice and point to improvements that can be made or alternative (better) forms of provision. Some tiered rating systems encompass a wide range of practice, from fledgling interventions with little or no evaluation to established interventions with evidence of effectiveness from multiple rigorous trials. This can help commissioners or policymakers to see where locally delivered services sit on this spectrum, and reflect on whether they need to refocus current provision by introducing different interventions or improving those already in place. There have been efforts by some registries to aid this process. One involves conducting subject-specific evidence reviews in areas such as early learning and interparental conflict (Asmussen, Feinstein, Martin, & Chowdry, 2016; Harold, Acquah, Chowdry, & Sellers, 2016), plotting interventions used in regular practice[9] against registry standards and showing how they compare with better-evidenced (but often less widely disseminated) alternatives. Another approach, drawing on different standards, offers a roadmap for strengthening aspects of intervention specification and system readiness (Axford et al., 2013).

6.4. Critical thinking

There is anecdotal evidence that standards and registries have encouraged commissioners, managers and practitioners to reflect more critically about the interventions they are involved with – what they seek to do, how, and with what success. This is partly achieved by promoting the testing of intervention effectiveness, and demonstrating that some well-meaning and ostensibly sensible interventions are ineffective or even harmful. However, critical reflection is also nurtured through the process of developing a logic model or theory of change, a foundational requirement in many standards. This gives stakeholders a language for talking about why what they do should work. That said, not all registries report the intervention theory of change, and fewer still have standards for assessing its plausibility and resonance with best evidence – a clear area for improvement.

[9] As long as they have an evaluation.

6.5. Innovation

Whether standards and registries have promoted innovation in intervention development is unclear, although there are several reasons to think that they can do so. First, they provide pointers to the kinds of things that do and don't work in terms of improving outcomes. Second, evidence of how much, with whom and through which mechanisms interventions are effective usually points to elements needing further innovation and research. Third, by categorizing interventions (e.g., outcomes, target group, setting) and their strength of evidence, registries help with identifying gaps in knowledge and practice. Fourth, standards can underpin initiatives to fund low-tier but promising interventions with a view to supporting their improvement against standards.[10]

However, critics contend that registries stifle innovation because commissioners restrict funding to highly-rated programs. Whether this is true is debatable; as noted already, highly-rated interventions hardly dominate practice, while locally developed interventions – tested and untested – continue to be prevalent in child welfare, education and juvenile justice. Moreover, even if true, it is not entirely without merit. There is a strong case for investing in interventions with demonstrable effectiveness over those that are (i) untested but likely – given the evidence base – to be ineffective or harmful, or (ii) known to be ineffective or harmful. Registries can help discourage re-inventing the wheel, encourage genuine innovation and, with better organization of the evidence (e.g., drawing out elements of effective practice), support the improvement of existing interventions.

7. Consistency and efficiency

There are many brands of standards and registries. In one respect this is understandable and even desirable: the variety potentially caters for different audiences and needs. For instance, some registries are country-specific while others are pan-national, and some focus on one subject (e.g., substance use, education, crime) while others have a broader remit (e.g., early intervention, positive youth development). Another (perverse) driver of diversity is the incentive for organizations to create their own standards and websites to build brand and create intellectual property.

Unfortunately, the result is confusing and inefficient, and not only because it isn't obvious to users which registry to use. Programs do not necessarily appear consistently where they might reasonably be expected. For example, one study found that 79% of programs listed in registries of individual programs appear on only one such registry, despite being eligible for an average of 5.6 additional registries (Means et al., 2015). There is also duplication: some programs appear in more than one registry, creating redundancy where they agree and uncertainty or puzzlement for decision-makers where ratings differ (Burkhardt et al., 2015; Zack et al., 2019). It is also not uncommon for an intervention to receive a high rating on one registry and an apparently lower one elsewhere (the diversity of rating systems makes comparison difficult, at least intuitively). For instance, Means et al. (2015) examined a random sample of 100 programs assessed by more than one registry and found that 53% received different classifications across organizations.

There are several reasons for these inconsistencies (Means et al., 2015; Fagan & Buchanan, 2016; Zack et al., 2019). First, registries use different processes and criteria to select studies that inform the rating. Key differences include: conducting a comprehensive search or relying on submitted materials; stipulating peer-reviewed literature or including grey literature; treating an adapted version of a program as a

[10] In the UK, these include Realising Ambition (funded by Big Lottery) and the Education Endowment Foundation and Youth Endowment Fund What Works Centres.

new entity or badging it under the old version; using geographic location of studies as a filter (or not); and only including comparison group studies or permitting a wider range of evaluation designs. The net effect is that registries can review different evidence bases for the same intervention. Second, the criteria used to rate study quality vary between registries. Common differences relate to the treatment of outcome measures' reliability and validity, the use of intent-to-treat analysis, sample representativeness, the tolerance of attrition and baseline imbalances, and the quality of analysis. Third, the rating of effectiveness is inconsistent. Some registries focus only outcomes of interest to a specific agency or government department, overlooking effects on other outcomes. Discrepancies higher up rating scales include whether value is assigned to evidence of sustained effects post-intervention or the independent replication of effects. Fourth, as noted, registries diverge over the need for interventions to be "dissemination ready". In short, each registry assesses and rates programs in a particularistic fashion.

The resulting danger of confused but time-poor commissioners or policymakers losing faith in ratings and registries indicates the need for some degree of consolidation or benchmarking. This should not be too difficult, given that most standards can arguably be traced back to several key sources, such as the Maryland Scientific Methods Scale (Sherman et al., 1997), the CONSORT statement (Moher et al., 2010) and the Society for Prevention Research standards (Flay et al., 2005; Gottfredson et al., 2015). There have been attempts to do this: the Annie E. Casey Foundation's *Evidence2Success* project convened the guardians of four well-known US registries[11] to create a common standard; the UK Alliance for Useful Evidence undertook a comparative analysis of standards (Puttick, 2018); and the Pew Results First registry collates ratings from nine national registries in the US. Unfortunately, these have not halted the proliferation of seemingly new but ultimately derivative approaches: more action is needed. A moratorium on creating new standards that cannot demonstrate significant added value would be a start, as would a resolve to plug acknowledged gaps in standards (e.g., on logic models). Increased coordination across registries would reduce uncertainty and redundancy while also releasing capacity to address other issues identified here, but critically can only happen if funders and host organizations commit to this goal.

8. Ethics

Ethical issues regarding registries are rarely discussed. On the positive side, registries (and similar knowledge mobilization tools) exist to help make judgments about evidence more transparent, systematic, efficient, and open to debate (Gough, 2021). In turn, by providing accurate and complete information about options, including evidence and costs, registries can help communities and stakeholders to make informed choices about intervention adoption or scale-up.

However, several ethical concerns warrant attention, starting with the *modus operandi* of the expert panels that commonly apply criteria and determine intervention ratings. Care is needed to deal appropriately with conflicts of interest, particularly when panel members stand to gain reputationally or financially from the rating assigned. This is not restricted only to interventions they have helped develop or evaluate directly. Normally, relevant panel members are not permitted to participate in discussions about the intervention in question, but it pays to be alert to more subtle forms of influence.

Next, the standards or criteria used to rate interventions have an ethical angle. For instance, it is unusual to consider explicitly whether an intervention reduces social and health inequities along axes of potential disadvantage, such as place of residence (e.g., urban/rural), gender/sex, SES and race/ethnicity/culture/language. This limits the usefulness of the assessment for policymakers, not least because it leaves open the possibility of unintentional intervention-generated inequities. Recommendations for how systematic reviews can address this issue (Welch, Petkovic, Jull, Hartling, Klassen, & Kristjansson, 2021) arguably apply to registries also, for example looking at whether potentially disadvantaged populations achieve the same improvement in outcomes (in absolute and relative terms) and considering less tangible (but important) outcomes for participants, such as inconvenience, burden (out-of-pocket costs, travel time) and stigma. This could help avoid registries assigning high ratings to interventions that widen – or fail to narrow – inequities, or at least provide decision-makers with a fuller picture of the evidence. Of course, deficits in relevant information may lie in the primary studies but registries can still report as much.

A further ethical challenge relates to the dissemination of interventions that receive registry endorsement. For example, there is a risk of promoting interventions that are effective when delivered with fidelity but to communities that lack the resources to achieve this. This may lead to negligible or iatrogenic effects. There is also legitimate apprehension about commercialization, a key strategy for supporting the dissemination of EBIs. Specifically, it may restrict access for disadvantaged groups if charges for materials and technical assistance exceed school or community group budgets, or if there are copyright or intellectual property restrictions. Commercialization can also create marketing pressures that lead to findings being overstated (Leadbeater et al., 2018).

9. Where next for registries?

Decision-makers need to consider a range of factors when choosing how to invest in interventions to improve child and youth psychosocial outcomes, including ethics, equity, politics, pragmatics, context, value for money and scientific evidence. In order to maximize the usefulness of evidence, it must be presented in a format that is easy to interpret and apply. This requires good governance based on principles such as transparency, contestability, and integrity (OECD, 2020). In this light, the emergence of "what works" registries and associated standards of evidence is a positive development. Previously, it was difficult for policymakers and commissioners to find relevant studies, let alone appraise their quality and findings. It was even tougher to compare interventions and discriminate between those found to be effective, those known to be harmful and those with no evidence of impact. By some accounts, everything "worked", which was manifestly untrue and a by-product of an accountability culture that requires service providers to demonstrate value to funders. Now it is harder to get away with unfounded claims about effectiveness, or to defend using any intervention because there is no good evidence in the field. Registries can serve as honest brokers by informing decisions from an ostensibly neutral standpoint while being open to scrutiny and questioning; for example, developers can usually contest ratings they consider unreasonable.

However, there is clearly scope to improve their impact on service commissioning. Aside from encouraging greater *coordination* across registries, strategies advanced in this article include attending to hitherto neglected issues in registry *content*, such as transportability to new contexts and first-hand implementation experiences, and ensuring that potential registry users are involved in the *design* of registry interfaces and functionality. Parallel efforts are need to improve the *application* of registry content. These entail raising awareness of registries among intended users and promoting their intelligent use. Critically, ratings should not be used deterministically; highly-rated interventions should not be adopted unthinkingly, just as those with equivocal or no evidence of impact should not automatically be discarded. Fostering this more considered approach will require understanding decision-makers' needs and offering tailored training and support, whether in navigating and interpreting registry content (e.g., via pop-up videos, text explainers, online tutorials, instant messaging) (Karre et al., 2017) or making sense of evidence more generally.

There is also a strong case for using registries primarily in the context

[11] Blueprints for Healthy Youth Development; Communities that Care; Best Evidence Encyclopedia; and Child Trends LINKS.

of structured methods for planning and supporting community-wide prevention efforts, such as *Communities that Care* (Fagan, Hawkins, Farrington, & Catalano, 2018). These help stakeholders to select interventions that best address the local risk and protective factor profile and to implement those interventions well. They also engage local communities, thereby broadening the range of views on the evidence and increasing buy-in among potential service users. This is valuable because public representation is a core principle in the good governance of evidence (OECD, 2020). In short, registries may need to be embedded in systems if they are to realize their potential.

More radical developments are needed, however, to achieve step changes in making prevention and early intervention more evidence-based. The first set concern the evaluation methods recognized and encouraged by registries. Intervention development is often more organic than standards and registry rating systems imply, meaning that improvement is not necessarily driven by the summative evaluation approaches in which most standards are grounded. Further, the vast majority of interventions in practice are unlikely to be trialled, whether because it is impossible or unnecessary. Additionally, interventions are essentially "events in systems", meaning that it is insufficient to consider them in isolation or restrict assessment of impact to a few tightly-specified outcomes. This is particularly pertinent given growing understanding of the complexities of youth psychosocial problems and efforts to prevent them.

Evaluation practice therefore needs a more expansive repertoire of (non-trial) methods for assessing intervention impact. These methods involve either mimicking control groups or exploring whether mechanisms articulated in the theory of change have been actualized. A more formative or developmental approach to assist managers and practitioners with making data-informed decisions may also be suitable. Further, there is a need for more system-based evaluation. The standards that underpin registry rating systems need refocusing in order to acknowledge these shifts in emphasis.

A second group of changes to how registries operate relate to how "intervention" is conceived. The drawbacks of programs identified earlier indicate the need for a more mixed economy. At the simplest level, other forms of intervention with arguably greater potential for impact, notably policies and system reform approaches, need recognition. Currently, they tend to be overlooked by registries because they cannot easily be evaluated using the methods prized by most standards of evidence. A step on from this is to use accumulated evidence from program evaluations for purposes besides promoting said programs. A more *granular* approach, for instance, entails identifying and rating common elements of effective interventions; conditional on sufficient attention to evidential strength, feasibility and implementation issues, these can be used to build better programs, improve existing programs or, alternatively, inform regular practice through training and education (McKaskill et al., 2021). A more *global* turn involves distilling underlying models for clusters of similar interventions, and using these to inform evidence-based practice. In both cases, programs are viewed less as branded products to lift off the shelf and drop into new contexts and more as artifacts created during a process of generating knowledge (about what works, with whom, and how) to inform practice. Some registries do this to a degree, combining evidence on discrete interventions with insights on what, collectively, it implies for practice.

A third set of developments involves rethinking the impact pathways implicit in the registry approach. The prevailing orthodoxy in prevention and early intervention for improving outcomes, namely to develop programs, demonstrate their effectiveness in trials, and ultimately scale those that are successful, appears increasingly untenable, at least in isolation. A refocused approach would pay more attention to improving services as usual, hence the need to recognize alternative forms of intervention and evaluation. However, capitalizing on such an approach necessarily means employing methods to help change practice (e.g., Lipsey, Howell, Kelly, Chapman, & Carver, 2010), the challenges of which should not be underestimated (Liberman & Hussemann, 2016). An implementation science lens can help here. This emphasizes co-creating services with practitioners, making accommodations to context to optimize chances of adoption, and supplementing access to knowledge with ongoing technical assistance to support the behavior changes necessary for sustained effectiveness (Ghate, 2016). There is also a strong argument for embedding evidence and methods for its utilization into qualifying and continuing professional training for prevention and early intervention practitioners.

There are no easy answers to the issues raised in this article, of course, and there will be challenges in implementing the changes advocated here. Resource constraints need to be taken into account, as managing a registry is labour intensive, often requiring trade-offs between rigor and the time and resources it takes to complete reviews and website maintenance tasks (Burkhardt et al., 2015). Registries also face a perennial tension between information overload and oversimplification. Moreover, responsibility for making some of the changes advocated lies with others besides registry curators.

There is also much that is unknown about registries, pointing to the need for research by independent investigators. First, there is some research about the impact of registries and standards but it is limited and we need stronger answers to important questions. How widely are registries used, by whom, and what factors affect this? What effect does this have on how decision-makers and other users think and act, particularly regarding investment in EBIs? Are there adverse effects? Second, it is necessary to test the effect of innovations in registry content, design and application suggested here. For instance, does it affect registry use and impact if there is more information on websites about implementation experiences, or if stakeholders help design registries, or if registry users are supported in interpreting evidence? Third, there is a question about the extent to which registries and standards can support alternative models of intervention development and evaluation. Consideration needs to be given to their fit in a new evidence landscape in which pathways to impact are less through scaling-up discrete programs and more through alternative mechanisms.

10. Conclusion

Registries are a valuable addition to efforts to help services improve child and youth psychosocial outcomes. Immediate priorities for strengthening the existing offer include improving content, functionality and design, and supporting decision-makers with interpreting and applying registry content (accepting that this is not the responsibility of registries alone). These might be facilitated by greater coordination between registries (inasmuch as this is permitted by their governance), which in turn could help to improve consistency and perceived trustworthiness. However, more ambitious changes are needed to respond to key critiques (realist, system-orientated, implementation science) of the general approach and thereby increase the chances of impact. They include: saying more about what works for whom, when, in what context; recognizing a broader set of intervention types and evaluation methods; presenting information in a way that supports changing regular practice alongside the implementation of specific programs; and situating registries overtly in the context of evidence-informed strategies to support research utilization. Since registries exist within a broader evidence ecosystem, these changes necessarily involve a wider group of stakeholders besides those responsible for developing and maintaining registries, including funders, government, intermediaries, intervention developers and evaluators.

CRediT authorship contribution statement

Nick Axford: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. **Louise Morpeth:** Writing – review & editing. **Gretchen Bjornstad:** Writing – review & editing. **Tim Hobbs:** Writing – review & editing. **Vashti Berry:** Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: 'Two authors are involved with assessing programs for the Early Intervention Foundation (EIF) Guidebook (NA, VB) and the Xchange database of the European Monitoring Centre for Drugs and Drug Addiction (NA). NA is a member of the EIF and Xchange Evidence Panels. The other authors declare that they have no conflict of interest'.

Acknowledgements

We are grateful to Daniel Acquah, Gregor Burkhart, Frederick Groeger-Roth, Jack Martin and Tom McBride for helpful comments on a draft of this article. The time of Nick Axford, Vashti Berry and Gretchen Bjornstad is supported by the National Institute for Health Research Applied Research Collaboration South West Peninsula. The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.childyouth.2022.106469.

References

Adler, J. R., & Coulson, M. (2016). *The Justice Data Lab: Synthesis and Review of Findings*. London: Middlesex University.

Asmussen, K., Feinstein, L., Martin, J., & Chowdry, H. (2016). *Foundations for Life: What Works to Support Parent-Child Interaction in the Early Years*. London: Early Intervention Foundation.

Asmussen, K., Brims, L., & McBride, T. (2019). *10 Steps for Evaluation Success*. London: Early Intervention Foundation.

Axford, N., Berry, V., Blower, S., Little, M., Hobbs, T., & Sodha, S. (2013). *Design & Refine: Developing Effective Interventions for Children and Young People*. Dartington: The Social Research Unit.

Axford, N., Berry, V., Lloyd, J., Hobbs, T., & Wyatt, K. (2020). Promoting learning from null or negative results in prevention science trials. *Prevention Science*. https://doi.org/10.1007/s11121-020-01140-4

Baldus, C., Thomsen, M., Sack, P.-M., Bröning, S., Arnaud, N., Daubmann, A., et al. (2016). Evaluation of a German version of the Strengthening Families Programme 10–14: A randomised controlled trial. *European Journal of Public Health, 6* (December), 953–959.

Berry, V., Axford, N., Blower, S., Taylor, R. S., Edwards, R. T., Tobin, K., … Bywater, T. (2016). The effectiveness and micro-costing analysis of a universal, school-based, social-emotional learning programme in the UK: a cluster-randomised controlled trial. *School Mental Health, 8*(2), 238–256.

Brick, C., Freeman, A. L. J., Wooding, S., Skylark, W. J., Marteau, T. M., & Spiegelhalter, D. J. (2018). Winners and losers: Communicating the potential impacts of policies. *Palgrave Communications, 4*, 69.

Brooks, J. L., Boulay, B. A. & Maynard, R. A. (2019). *Empowering practitioners to drive the evidence train: Building the next generation of evidence*. Available at: https://www.projectevident.org/updates/2019/6/4/empowering-practitioners-to-drive-the-evidence-train-building-the-next-generation-of-evidence (Accessed 21st August 2020).

Buckley, P. R., Fagan, A. A., Pampel, F. C., & Hill, K. G. (2020). Making evidence-based interventions relevant for users: A comparison of requirements for dissemination readiness across program registries. *Evaluation Review, 44*(1), 51–83.

Burkhardt, J. T., Schröter, D. C., Magura, S., Means, S. N., & Coryn, C. L. S. (2015). An overview of evidence-based program registers (EBPRs) for behavioural health. *Evaluation and Program Planning, 48*(February), 92–99.

Burkhart, G., Axford, N., Sonthalia, S., Foxcroft, D., Faggiano, F., & De Kock, C. (2019). Why do flagship evidence-based programmes from the US run aground in Europe, and how should online repositories of programmes deal with this?. *European Society of Prevention Research 10th Annual Conference, Gent, Belgium, 9th September*.

Burton, R., Henn, C., Lavoie, D., O'Connor, R., Perkins, C., Sweeney, K., et al. (2017). A rapid review of the effectiveness and cost-effectiveness of alcohol control policies: An English perspective. *The Lancet, 389*, 1558–1580.

Cartwright, N., & Hardie, J. (2012). *Evidence-based Policy: A Practical Guide to Doing it Better*. Oxford: Oxford University Press.

Chen, H. T. (2015). *Practical Program Evaluation: Theory-driven Evaluation*. Thousand Oaks: Sage.

Egan, M., McGill, E., Penney, T., Anderson de Cuevas, R., Er, V., Orton, L., et al. (2019). *NIHR SPHR Guidance on Systems Approaches to Local Public Health Evaluation. Part 2: What to Consider When Planning a Systems Evaluation*. London: National Institute for Health Research School for Public Health Research.

Eisner, M. (2009). No effects in independent prevention trials: Can we reject the cynical view? *Journal of Experimental Criminology, 5*(2), 163–183.

Embry, D. D., & Biglan, A. (2008). Evidence-based kernels: Fundamental units of behavioral influence. *Clinical Child and Family Psychology Review, 11*(3), 75–113.

Fagan, A. A., & Buchanan, M. (2016). What works in crime prevention? Comparison and critical review of three crime prevention registries. *Criminology & Public Policy, 15* (3), 617–649.

Fagan, A. A., Hawkins, J. D., Farrington, D. P., & Catalano, R. F. (2018). *Communities that Care: Building Community Engagement and Capacity to Prevent Youth Behavior Problems*. Oxford Scholarship Online.

Fagan, A. A., Bumbarger, B. K., Barth, R. P., Bradshaw, C. P., Cooper, B. R., Supplee, L. H., et al. (2019). Scaling up evidence-based interventions in US public systems to prevent behavioral health problems: Challenges and opportunities. *Prevention Science, 20*(8), 1147–1168.

Faggiano, F., Allara, E., Giannotta, F., Molinar, R., Sumnall, H., Wiers, R., et al. (2014). Europe needs a central, transparent, and evidence-based approval process for behavioural interventions. *PLoS Medicine, 11*(10), Article e1001740.

Flay, B. R., Biglan, A. S., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., et al. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6*(3), 151–175.

Fletcher, A., Jamal, F., Moore, G., Evans, R. E., Murphy, S., & Bonell, C. (2016). Realist complex intervention science: Applying realist principles across all phases of the Medical Research Council framework for developing and evaluating complex interventions. *Evaluation, 22*(3), 286–303.

Fonagy, P., Butler, S., Cottrell, D., Scott, S., Pilling, S., Eisler, I., et al. (2018). Multisystemic therapy versus management as usual in the treatment of adolescent antisocial behaviour (START): A pragmatic, randomised controlled, superiority trial. *Lancet Psychiatry, 5*(2), 119–133.

Ford, T., Hutchings, J., Bywater, T., Goodman, A., & Goodman, R. (2009). Strengths and Difficulties Questionnaire Added Value Scores: Evaluating effectiveness in child mental health interventions. *British Journal of Psychiatry, 194*(6), 552–558.

Furlong, M., McGilloway, S., Bywater, T., Hutchings, J., Smith, S. M., & Donnelly, M. (2012). Behavioural and cognitive-behavioural group-based parenting programmes for early-onset conduct problems in children aged 3 to 12 years. *Cochrane Database of Systematic Reviews, 15*(2), CD008225.

Gennetian, L. A., Tamis-Lemonda, C. S., & Frank, M. C. (2020). Advancing transparency and openness in child development research: Opportunities. *Child Development Perspectives, 14*(1), 3–8.

Ghate, D. (2016). From programs to systems: Deploying implementation science and practice for sustained real world effectiveness in services for children and families. *Journal of Clinical Child & Adolescent Psychology, 45*(6), 812–826.

Goff, J., Hall, J., Sylva, K., Smith, T., Smith, G., Eisenstadt, N., Sammons, P., Evangelou, M., Smees, R., & Chu, K. (2013). *Evaluation of Children's Centres in England (ECCE) - Strand 3: Delivery of Family Services by Children's Centres*. Research Report DFE-RR297. London: Department for Education.

Gorard, S., Griffin, N., & See, B. H. (2019). *How Can We Get Educators to Use Research Evidence?* LULU Press.

Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., et al. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science. *Prevention Science, 16*(6), 893–926.

Gough, D. (2021). Appraising evidence claims. *Review of Research in Education, 45*, 1–26.

Gough, D., Maidment, C., & Sharples, J. (2018). *UK What Works Centres: Aims, Methods and Contexts*. London: PPI-Centre, Social Science Research Unit, UCL Institute of Education.

Gough, D., & White, H. (2018). *Evidence Standards and Evidence Claims in Web Based Research Portals*. London: Centre for Homelessness Impact.

Greenberg, M. T., & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness, 10*(1), 40–67.

Harold, G., Acquah, D., Chowdry, H., & Sellers, R. (2016). *What Works to Enhance Interparental Relationships and Improve Outcomes for Children?* London: Early Intervention Foundation.

Horne, C. S. (2017). Assessing and strengthening evidence-based program registries' usefulness for social service program replication and adaptation. *Evaluation Review, 41*(5), 407–435.

Humayun, S., Herlitz, L., Chesnokov, M., Doolan, M., Landau, S., & Scott, S. (2017). Randomized controlled trial of Functional Family Therapy for offending and antisocial behavior in UK youth. *Journal of Child Psychology and Psychiatry, 58*(9), 1023–1032.

Husk, K., Blockley, K., Lovell, R., Bethel, A., Lang, I., Byng, R., et al. (2020). What approaches to social prescribing work, for whom, and in what circumstances? A realist review. *Health & Social Care in the Community, 28*(2), 309–324.

Indig, D., Lee, K., Grunseit, A., Milat, A., & Bauman, A. (2018). Pathways for scaling up public health interventions. *BMC Public Health, 18*, 68.

James, A. C., James, G., Cowdrey, F. A., Soler, A., & Choke, A. (2013). Cognitive behavioural therapy for anxiety disorders in children and adolescents. *Cochrane Database of Systematic Reviews, 3*(6), CD004690.

Johnson, S. D., Tilley, N., & Bowers, K. J. (2015). Introducing EMMIE: An evidence rating scale to encourage mixed-method crime prevention synthesis reviews. *Journal of Experimental Criminology, 11*, 459–473.

Karre, J. K., Perkins, D. F., Aronson, K. R., DiNallo, J., Kyler, S. J., Olson, J., et al. (2017). A continuum of evidence on evidence-based programs: A new resource for use in military social service delivery. *Military Behavioral Health, 5*(4), 346–355.

Kemp, F., Ohlson, C., Raja, A., Morpeth, L., & Axford, N. (2018). Fund-mapping: the investment of public resources in the well-being of children and young people in Northern Ireland. *Child Care in Practice, 24*(4), 335–350.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., et al. (2016). Badges to acknowledge open practices: A simple, low-cost effective method for increasing transparency. *PLOS Biology, 14*(5), Article e1002456.

Kreuter, M. W., & Bernhardt, J. M. (2009). Reframing the dissemination challenge: A marketing and distribution perspective. *American Journal of Public Health, 99*(12), 2123–2127.

Knox, G., Hill, C., & Berlin, G. (2018). Can evidence-based policy ameliorate the nation's social problems? *Annals of the American Academy of Political and Social Science, 678*(1), 166–179.

Leadbeater, B. J., Dishion, T., Sandler, I., Bradshaw, C. P., Dodge, K., Gottfredson, D., et al. (2018). Ethical challenges in promoting the implementation of preventive interventions: Report of the SPR Task Force. *Prevention Science, 19*(7), 853–865.

Leijten, P., Gardner, F., Melendez-Torres, G. J., Van Aar, J., Hutchings, J., Schulz, S., et al. (2019). Meta-analyses: Key parenting program components for disruptive child behavior. *Journal of the American Academy of Child & Adolescent Psychiatry, 58*(2), 180–190.

Lemire, S., Christie, C. A., & Inkelas, M. (2017). The methods and tools of improvement science. In C. A. Christie, M. Inkelas, & S. Lemire (Eds.), *Improvement science in evaluation: methods and uses, 153* pp. 23–33). New Directions for Evaluation.

Liberman, A., & Hussemann, J. (2016). *Implementing the SPEP™: Lessons from Demonstration Sites in OJJDP's Juvenile Justice Reform and Reinvestment Initiative*. Washington DC: Urban Institute.

Lipsey, M. (2020). Revisited: Effective use of the large body of research on the effectiveness of programs for juvenile offenders and the failure of the model programs approach. *Criminology & Public Policy, 19*, 1329–1345.

Lipsey, M., Howell, J. C., Kelly, M. R., Chapman, G., & Carver, D. (2010). *Improving the Effectiveness of Juvenile Justice Programs: A New Perspective on Evidence-based Practice*. Washington DC: Center for Juvenile Justice Reform, Georgetown University.

Lortie-Forgues, H., Na Sio, U., & Inglis, M. (2021). How should educational effects be communicated to teachers? *Educational Researcher, 50*(6), 345–354.

Maranda, M. J., Magura, S., Gugerty, R., Lee, M. J., Landsverk, J. A., Rolls-Reutz, J., et al. (2021). State behavioral health agency website references to evidence-based program registers. *Evaluation and Program Planning, 85*, Article 101906.

Martin, J., McBride, T., Brims, L., Doubell, L., Pote, I., & Clarke, A. (2018). *Evaluating Early Intervention Programmes: Six Common Pitfalls, and How to Avoid Them*. London: Early Intervention Foundation.

Means, S. N., Magura, S., Burkhardt, J. T., Schröter, D. C., & Coryn, C. L. S. (2015). Comparing rating paradigms for evidence-based program registers in behavioural health: Evidentiary criteria and implications for assessing programs. *Evaluation and Program Planning, 48*(February), 100–116.

McKaskill, M., Axford, N., Hobbs, T., McNeil, B., Freeman, L., & Lily, R. (2021). *Learning and Adapting to Support Young People During the COVID-19 Pandemic: How a Core Components Approach Can Help*. London: Youth Endowment Fund.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., et al. (2013). The behaviour change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine, 41*(6), 81–95.

Mihalic, S. F., & Elliott, D. S. (2015). Evidence-based programs registry: Blueprints for Healthy Youth Development. *Evaluation and Program Planning, 48*(February), 124–131.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., et al. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, Article c869.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2015). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *Journal of Clinical Epidemiology, 70*(February), 191–199.

Neuhoff, A., Axworthy, S., Glazer, S., & Berfond, D. (2015). *The What Works Marketplace: Helping Leaders Use Evidence to Make Smarter Choices*. Boston: New York and San Francisco, The Bridgespan Group.

Nilsen, P. (2020). Overview of theories, models and frameworks in implementation science. In P. Nilsen, & S. Birken (Eds.), *Handbook on Implementation Science* (pp. 8–31). Cheltenham: Edward Elgar.

OECD. (2020). *Mobilising Evidence for Good Governance: Taking Stock of Principles and Standards for Policy Design, Implementation and Evaluation*. Paris: OECD Publishing.

Oliver, K., & Boaz, A. (2019). Transforming evidence for policy and practice: Creating space for new conversations. *Palgrave Communications, 5*, 60.

Patton, M. Q. (2010). *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York, NY: Guilford Press.

Piazza, R., Corry, D., Noble, J., & Bagwell, S. (2019). *Data Labs: A New Approach to Impact Evaluation*. London: New Philanthropy Capital.

Puttick, R. (2018). *Mapping the Standards of Evidence Used in UK Social Policy*. London: Alliance for Useful Evidence.

Rutter, H., Savona, N., Glonti, K., Bibby, J., Cummins, S., Finegood, D. T., et al. (2017). The need for a complex systems model of evidence for public health. *Lancet, 390*, 2602–2604.

Sherman, L. W., Gottfredson, D., Mackenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing Crime: What Works, What Doesn't, What's Promising*. Washington DC: US Department of Justice.

Sigfusdottir, I. D., Kristjansson, A. L., Gudmundsdottir, M. L., & Allegrante, J. P. (2011). Substance use prevention through school and community-based health promotion: A transdisciplinary approach from Iceland. *Global Health Promotion, 18*(3), 23–26.

Skärstrand, E., Sundell, K., & Andréasson, S. (2013). Evaluation of a Swedish version of the Strengthening Families Programme. *European Journal of Public Health, 24*(4), 578–584.

Steeger, C. M., Buckley, P. R., Pampel, F. C., Gust, C. J., & Hill, K. G. (2021). Common methodological problems in randomized controlled trials of preventive interventions. *Prevention Science, 22*(8), 1159–1172.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the Range of Designs and Methods for Impact Evaluation*. Working Paper 38. London: Department for International Development.

Sundell, K., Hansson, K., Löfholm, C., Olsson, T., Gustle, L.-H., & Kadesjö, C. (2008). The transportability of Multisystemic Therapy to Sweden: Short-term results from a randomized trial of conduct-disordered youths. *Journal of Family Psychology, 22*(4), 550–560.

SUPERU (Social Policy Evaluation and Research Unit). (2016). *Standards of Evidence for Understanding What works: International Experiences and Prospects for Aotearoa New Zealand*. Wellington: NZ SUPERU.

Tanner-Smith, E. E., Durlak, J. A., & Marx, R. A. (2018). Empirically based mean effect size distributions for universal prevention programs targeting school-aged youth: A review of meta-analyses. *Prevention Science, 19*(8), 1091–1101.

Teager, W., Fox, S., & Stafford, N. (2019). *How Australia Can Invest Early and Return More: A New Look at the $15b Cost and Opportunity*. Australia: Early Intervention Foundation, The Front Project, and CoLab at the Telethon Kids Institute.

Welch, V. A., Petkovic, J., Jull, J., Hartling, L., Klassen, T., Kristjansson, E. et al. (2021) Equity and specific populations. In: Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J. & Welch, V. A. (Eds) *Cochrane Handbook for Systematic Reviews of Interventions* version 6.2 (updated February 2021). Available from www.training.cochrane.org/handbook.

White, H., & Phillips, D. (2012). *Addressing Attribution of Cause and Effect in Small N Impact Evaluations: Towards an Integrated Framework*. Working Paper 15. New Delhi: International Initiative for Impact Evaluation.

Zack, M. K., Karre, J. K., Olson, J., & Perkins, D. F. (2019). Similarities and differences in program registers: A case study. *Evaluation and Program Planning, 76*(October), Article 101676.