The International Journal of Psychosocial Rehabilitation
Volume 20, Number 1
July 2015– June 2016

Assessing the Suitability of Clinical Programs for Implementation

Tom Bartholomew
Joe Birkmann

Rutgers University

Citation:
Bartholomew T & Birkman J. (2016)
Assessing the Suitability of Clinical Programs for Implementation
International Journal of Psychosocial Rehabilitation. Vol 20 (1) 112-119

Abstract

Mental health organizations looking to implement new clinical programs are faced with an ever-increasing number of options to choose from. Determining which program(s) to implement is often based on a fondness for "pet programs" and factors such as the availability of vender packages that may have little bearing on the appropriateness of a given program. A nine-dimension rubric is proposed as a way of measuring a program's suitability for implementation. This rubric involves an assessment of a program's: objective(s), efficacy, generalizability, cost benefit profile, opportunity cost(s), fidelity measurement, outcome assessment, feasibility, and three factors related to implementation. These dimensions of suitability are presented as score-able criteria to offer organizations a means to compare and contrast various clinical programs. Programs are scored, ideally first by venders or program advocates, then individually by those charged with making a decision about implementation. Lastly, consensus is sought on scores across the nine-dimension rubric using the measurable anchors. Limitations of this approach are discussed. Future work in this area is recommended.

The authors have no financial disclosures or conflicts to report.

Introduction:

Mental health organizations are charged with evaluating and implementing effective clinical programs for the benefit of their clients. The suitability of programs for implementation is often a complex, context dependent match between programs, organizational characteristics and client's needs (Blase & Fixsen, 2013; Fixsen et al., 2005). Organizations must actively manage their "clinical formulary" in order to maximize the effective use of limited resources. Historically organizational leadership may have exercised a deferential attitude toward professional degrees and titles, opting for what was essentially the product of a "private practice" model of service delivery. In this approach practitioners could exercise their professional judgment regarding the interventions that they chose to use (Drake , Merrens & Lynde, 2005) without necessarily considering the organization's needs. This can lead to what Carol Mowbrey (2003) has called "black box" outcomes in which a professional does something which may or may not have an effect but what was done remains unknown. The consequence

of this lack of transparency and accountability is that outcomes can be unreliable and clients can and do suffer unnecessary treatment failures. The PORT studies (Leman et al., 2004: Dixon et al., 2010) demonstrate that the majority of mental health organizations in the United States are failing to use practices that have demonstrable effectiveness in favor of "private practice" approaches that are largely based, not on evidence, but on practitioner preferences. The need for accountability in the provision of programming is becoming more important as payers demand an evidence based clinical formulary in order provide reimbursement (Fox, 2005). In order to assist organizations and practitioners in assessing the suitability of programs for implementation a score-able nine-dimension rubric has been developed.

The nine-dimension rubric (see appendix 1) allows for the assessment of a clinical program's suitability for implementation using an anchored five point scale (1 to 5), allowing for a program scoring range of between 9 and 45. Higher scores indicate greater program suitability. The intent of the rubric is not to achieve an absolute measure of program suitability. Instead the goal is to promote the development of consensus across multiple raters (see scoring grid at the bottom of appendix 1). In this way important aspects of the decision-making regarding program suitability require a focused and informed discussion focused on achieving a collective scoring decision. Multiple staff can compare multiple programs across the same nine dimensions. Another effective way to use the rubric is to have venders or program supporters score the rubric and to provide the necessary information required to score the rubric. In this way the "burden" of demonstrating the suitability of the program for a given organization at a given time falls on the proponent of the program. The information presented by the program's advocate is then confirmed or rejected by consensus of the larger agency group. The scoring rubric assesses a program's: objective(s), efficacy, generalizability, cost benefit profile, opportunity cost(s), fidelity measurement, outcome assessment, feasibility, and three factors related to implementation. An "other" dimension can be added and scored in the event that there are idiosyncratic factors that affect the suitability of a particular program and can leverage an otherwise low or high score.

The first item on the rubric scores the suitability of the objective(s) of a clinical program (Blase & Fixson, 2013). Agencies define their clinical mission to include who they are going to serve, in what way, for how long and by what approach. Aroma therapy may have demonstrable effectiveness for reducing behavioral disturbances in in patients with dementia (Smallwood et al., 2001), but it is unlikely to teach an individual with severe mental illness daily living skills. The question for this first rubric item involves whether the program being rated is consistent with the clinical mission, when considering the agency's current "clinical formulary". A clinical formulary can be understood much like a medication formulary, constituting the range of interventions or "programs" provided by an organization. The range of programs may include clinical strategies designed to meet the core aspects of particular conditions. Programs may also address more generic issues common to many clients, or programs may be primarily recreational. What is important is that an agency's clinical formulary is sufficient to meet the clinical mission of the agency. Aroma therapy may make good sense as an elective adjunct to other possibly more core strategies, or it may be considered a core strategy for use in accomplishing the agency's mission.

The second rubric item is efficacy. The idea of rating efficacy is to determine the strength of the causal inference that the program was the cause of a change in an important criterion variable (important outcome) in the face of other possible explanations for the result (Shadish, Cook & Campbell, 2009). Efficacy can sometimes, though this is not required, be quantified by calculating the effect size of an intervention. One of the most common calculations of effect size is Cohen's "d " (1977). This "effect size" calculation can be done by subtracting the mean of the dependent variable of the control group from the mean of the dependent variable of the experimental group and then dividing by the standard deviation of the control group. This calculation can be combined across many similar studies to form a very stable meta-analytic measure of the program's efficacy. Short of this, raters can simply rate the known level of evidence of the program from low to high. It is important to note that not all evidence is equal and raters should understand how to assess the strength of the evidence for a program (Guyatt et al., 2004).

Generalizability is the third rubric dimension and it involves the confidence that a program's efficacy referred to in rubric item #2 is transferable to the population and setting of interest. This can be assessed by determining if the research demonstrating the program' efficacy was done in a similar setting with a similar population under similar conditions.

The fourth dimension of the rubric is an estimate of the cost of the program compared to the probable benefits. This is understandably a subjective assessment and reliable only at the extremes of a very good or very poor cost benefit ratio. Some clinical programs such as aroma therapy could reach virtually all clients and be relatively inexpensive (Low Cost). This same program, however, may have only a moderate impact on the agencies clinical mission (Low Benefit). Another program, dialectical behavior therapy, may reach few clients and be very costly. This program may potentially have a large impact on the clinical mission of the agency (high cost / high benefit). The importance of this dimension is to achieve consensus about the relative cost to benefit ratio of various programs in a given setting in the face of limited resources.

The fifth dimension is opportunity cost. This is related to the cost/ benefit dimension but focuses on the program's value in relation to other possibly more efficacious programs. This rating assumes that the implementation of a given program is mutually exclusive of the implementation of another program and that the benefit(s) of the excluded program will not be realized. This rating requires that the rater(s) have an understanding of the range of similar available programs and their relative benefits. For this reason this dimension, while a critical consideration, is likely the most difficult and least reliable of all the item scores.

The sixth rubric score involves the availability of a fidelity or treatment adherence scale for the program being considered. Fidelity or treatment adherence involves the ability of practitioners and administrators to have objective feedback based on a clinical or programmatic "audit" of the degree that a program's critical ingredients have been implemented effectively. This can be accomplished by measuring the fidelity of the program implementation. Fidelity is the degree of adherence to the critical ingredients of a program model (Bond et al, 2000). Programs can fail to produce desired outcomes because of a failure to maintain fidelity to important aspects of the model that were assessed in rubric item #2 and that are thought to produce positive outcomes.

The seventh and related score on the rubric involves the ability to determine the outcome(s) of the intervention. Outcome assessment serves two purposes. The first is to be able to assess, in conjunction with a measure of fidelity if the program is effective at meeting its objectives. The second is to offer practitioners feedback about their efforts and the need to make corrections or adaptations in their approach.

The eighth rubric score involves the feasibility of an organization to implement a given program. Also know as organizational readiness, feasibility can be broken into structural and psychological readiness (Weiner, 2009). Structural readiness is a measure of whether the organization has sufficient resources, time, money, space, staff expertise, etc. The second level of readiness can be thought of as psychological readiness and involves the leadership and staff's level of motivation, support, and "buy in" to implementation of the program. There are many ways to measure organizational readiness in an organization. One simple way is to describe the program using the rubric and ask how staff and clients feel about implementing it. Organizational readiness is not a static state and can be increased or decreased depending on actions of agency leadership and local conditions. It is important to note that both forms of organizational readiness will vary according to the program being considered.

The ninth and last rubric item involves the plan for implementation of the program. Implementation necessarily involves more than training. Training programs that lack direct observation of staff behavior do not allow for staff to receive constructive feedback (Bandura, 1989, Kirkpatrick, 1979). This can result in staff believing they are skilled when they are not or can result in staff simply ignoring the training recommendations (Miller

& Mount 2001). The inclusion of consultation and supervision provide the best chance of the program being implemented to high fidelity and producing the desired outcomes. Without supervision, even programs that start out with high fidelity can suffer from a condition called practitioner drift in which the practitioner gradually reverts to a preferred, possibly ineffective, clinical approach (Bond et al, 2000).

Discussion

The assessment of the suitability of clinical programs for implementation is not an exact science nor should it be uninformed guesswork. Limited resources as well as the responsibility to provide effective interventions requires that conflicting priorities be reconciled. The implementation of pet projects or programs that are familiar to staff but are ineffective should not take precedence over programs that are likely to provide greater benefits (i.e. have a higher rubric score) in a given setting. The value in assessing and seeking consensus regarding the suitability of programs for implementation is that it allows for the triaging of an implementation decision. This enables agency staff and leadership to determine whether enough is known about the program to move it to the next phase of consideration. After the use of the rubric, it is recommended that a more detailed examination of implementation issues be undertaken. For a guideline to a more in-depth process see SAMHSA (2014).

Conclusion

The program assessment rubric (PAR) offers organizations a means of quickly assessing and comparing the suitability of programs for implementation in a given setting. The PAR is not intended as an absolute measure of program suitability. Instead, the intention of the PAR is to promote an examination of important aspects of program implementation and to allow the building of consensus across staff. A second purpose of the PAR is to arm staff with the ability to present dissenting opinions about program suitability in the face of pet projects or familiar programs that are unlikely to produce an agency's desired outcomes. Once programs have been vetted through consensus scoring using the PAR, it is recommended that additional work be done in order to prepare for implementation. Lastly the PAR can be used in a retrospective way to assess programs for exnovation (discontinuation) or to assess the reason(s) that programs failed or were abandoned. Program venders and sponsors can be asked to score the rubric and to provide the information and evidence needed to score the PAR. Limitations of the PAR begin with the lack of inter-rater reliability. The PAR is not intended as an absolute measure of program suitability, and as such, differences in scores of the rubric factors are expected and suggest the need for more information and or discussion in order to achieve consensus. The final decision about implementation is necessarily context dependent and may, in the end, be based on idiosyncratic factors. An additional limitation is the lack of weighting of the rubric items. Some rubric dimensions are likely to be more predictive of suitability than others. The value of rubric scoring may be in its simplicity. Future research into the assessment of program suitability for implementation should seek to determine which of the rubric dimensions are most predictive of successful implementation and under what conditions. Additional work in this area might also attempt to refine the behavioral anchors in an effort to improve inter-rater reliability.

References

Blase, K., Fixsen, D. (2013). Stages of Implementation Analysis: Where are we Now? National Implementation Research Network.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., Kim, H. W., & Bond. (2000). Measurement of fidelity in psychiatric rehabilitation. Mental Health Services Research, 2(2), 75–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11256719

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev). Lawrence Erlbaum Associates, Inc.

Dixon, L. B., Dickerson, F., Bellack, A. S., Bennett, M., Dickinson, D., Goldberg, W., Goldberg, R. W. (2010). The 2009 schizophrenia PORT psychosocial treatment recommendations and summary statements. Schizophrenia Bulletin, 36(1), 48–70. doi:10.1093/schbul/sbp115

Drake, R. E., Goldman, H. H., Leff, H. S., Lehman, a F., Dixon, L., Mueser, K. T., & Torrey, W. C. (2001). Implementing evidence-based practices in routine mental health service settings. Psychiatric Services (Washington, D.C.), 52(2), 179–82. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11157115

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M. & Wallace, F. (2005). Implementation Research: A Synthesis of the Literature. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Fox, D. (2005). Evidence of evidence-based health policy: the politics of systematic reviews in coverage decisions. Health Affairs, 24(1), 114-122.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ (Clinical Research Ed.), 336(7650), 924–6. doi:10.1136/bmj.39489.470347.AD

Lehman, A. F., Kreyenbuhl, J., Buchanan, R. W., Dickerson, F. B., Dixon, L. B., Goldberg, R., … Steinwachs, D. M. (2004). The Schizophrenia Patient Outcomes Research Team (PORT): updated treatment recommendations 2003. Schizophrenia Bulletin, 30(2), 193–217. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15279040

Mowbray, C. (2003). Fidelity Criteria: Development, Measurement, and Validation. The American Journal of Evaluation, 24(3), 315–340. doi:10.1016/S1098-2140(03)00057-2

SAMHSA (2014). NREEP, SAMHSA, National Registry of Evidence-based Practices Programs and Practices. Retrieved from :http://www.nrepp.samhsa.gov/pdfs/Questions_To_Ask_Developers.pdf
Smallwood, J., Brown, R., Coulter, F., Irvine, E., & Copland, C. (2001). Aromatherapy and behaviour disturbances in dementia: a randomized controlled trial. International Journal of Geriatric Psychiatry, 16(10), 1010-1013.

Weiner, B. J. (2009). A theory of organizational readiness for change. Implementation Science : IS, 4, 67. doi:10.1186/1748-5908-4-67

Weiner, B. J., Amick, H., & Lee, S.-Y. D. (2008). Conceptualization and measurement of organizational readiness for change: a review of the literature in health services research and other fields. Medical care research and review : MCRR (Vol. 65, pp. 379–436). doi:10.1177/1077558708317802

Appendix 1 **Program Assessment Rubric** v2

This rubric is designed as a tool to assess the suitability of programs for implementation in a variety of health care settings. **Directions**: Score proposed programs across the 9 criteria below and sum the scores (9-45). Higher total scores indicate a better program match to local conditions. This rubric is meant to facilitate conversation and is not intended as an absolute measure of program suitability. Groups of raters should compare individual scores and try to achieve consensus. Idiosyncratic factors may have great importance and trump otherwise low or high scores.

Criteria	1= Low	2	3= Medium	4	5 = High
1. Objective of the Program: To what degree is the intervention consistent with the agencies core clinical mission	The program is not specifically focused on the core clinical mission of the organization		The program has some baring on the core clinical mission of the organization		The program is specifically focused on the core clinical mission of the organization
"Objective" involves whether the focus of the program is consistent with the agencies clinical mission and needs of its clients. Some practices may have all the elements of a good program but do not belong in the "clinical formulary" of a given agency. This can occur when there is a mismatch between a program's focus and the necessity for an agency to meet its client's primary needs.					
2. Efficacy: What is the ability of the intervention to produce desired changes (outcomes) and to what degree.	No studies available on the program's efficacy. (d=.3 or <)*		There is a moderate amount of evidence of efficacy of the program. (d= between .3 and .5)		There is a large amount of high quality evidence on the efficacy of the program (d=.5or >).
Efficacy (Internal Validity) involves the quality of the evidence showing that a program is effective. A standardized "effect size" can be used to compare programs. This is called Cohen's d*. If the mean score of the studies criterion variable is important (Ex. Beck depression inventory) then a large Cohen's d score means that most clients were helped a lot. **If this is unavailable one can score the program based on the amount of available evidence.** * Cohen's d can be calculated by subtracting the control group mean from the experimental group mean and dividing by the standard deviation of the control group. Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (rev).* Lawrence Erlbaum Associates, Inc.					
3. Generalizability: What is the intervention's effectiveness with the setting and population of interest.	No studies of conducted with the setting and population of interest.		Some studies exist for the setting and population of interest though it may be quasi- experimental or a single RCT*.		Multiple RCTs with the setting and population of interest.
Generalizability (external validity) is the extent that the results of a study (ies) can be generalized to other situations and to other people. For example, "inferences based on comparative psychotherapy studies often employ specific samples (e.g. volunteers, highly depressed, no comorbidity). If psychotherapy is found effective for these sample patients, will it also be effective for non-volunteers or the mildly depressed or patients with concurrent other disorders? Aronson, et al. (2007). Social psychology. (4 ed.). Toronto, ON. * Unlike an RCT, Quasi- experimental describes studies that do not randomly assign subjects to the experimental condition. RCT = Randomized Controlled Trial, one of the best forms of evidence of a program's effectiveness.					
4. Cost / benefit: How many clients will benefit (or how will the organization benefit) from the intervention compared to the cost of the intervention	Benefits will be low relative to the cost. (<33% of clients are appropriate)		There will be medium benefits relative to cost. (Between 33% and 66% of clients are appropriate)		Benefits will be high relative to the Cost (> 66% of clients are appropriate)
Cost / benefit involves the balance of the cost to the organization in resources compared to the clinical and organization benefits that could be gained from implementation. Resource intensive programs that benefit high resource utilizers may be have a low cost benefit ratio. Despite this a small number of Clients with unmet clinical needs may benefit a great deal from resource intensive interventions when no other options exist for them.					
5. Opportunity Cost: Does implementing this intervention preclude the implementation of	Implementation of this program will consume resources needed for other more		Implementation of this program may consume some resources needed for other more		Implementation of this program will not consume resources needed for other more

another, possibly more effective, intervention?	demonstrably effective interventions.		demonstrably effective interventions.		demonstrably effective interventions.
Opportunity cost involves the assessment of the best use of limited resources when the implementation of one program may preclude the use of another.						
6. Fidelity: The program has tools for determining the effectiveness of program implementation.	The program has no measure of program fidelity.		The program has an un-validated measure of fidelity.		The program has a validated measure of fidelity.
Programmatic fidelity scales measure the degree to which a program adheres to a program model. Failure to assess fidelity can result in an otherwise effective program failing to produce positive outcomes due to poor adherence to the model. Bond, et al. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research*, *2*(2), 75–87.						
7. Outcome Assessment: The program has a measure of outcomes.	The program has no measure of outcome.		The program has no recommended outcome measures but existing measures can be used.		The program provides a recommended process for determining outcomes.
Outcome measures assess change(s) in clients that may be caused by the intervention. Outcomes differ from outputs in that the former is a measure of the effectiveness of the program. Output is a measure that a program was provided regardless of its quality.						
8. Do-ability: The organization has sufficient readiness to implement the intervention as designed.	Structural and psychological readiness is low.		Structural and psychological readiness is medium.		Structural and psychological readiness is high.
Also called Organizational Readiness and has two parts, the first involves whether the organization has sufficient "structural readiness" (resources, money, time, space, staff expertise etc.) and the second is psychological readiness (stake holder support, motivation, enthusiasm and "buy in" for the intervention). Readiness is likely different for each project. Weiner, B. J. (2009). A theory of organizational readiness for change. *Implementation Science : IS*, *4*, 67.						
9. Implementation: Training, consultation and supervision is part of the proposal / program	Only training is available.		Training and consultation is available.		Training, consultation and supervision (with direct observation of practice) is available.
Successful implementation involves the transfer of knowledge, skills and abilities over time through training, direct observation of practice and feedback. Training alone is insufficient to change clinical practice. Fixsen et al., (2005). Implementation Research: A Synthesis of the Literature. Tampa, FL: University of South Florida, Mental Health Institute.						

Scoring: Write your rating (1-5) in the box under the criteria. When done, add the scores and put the total in the box to the far right. Note the program being rated below:	1. Objective of Pro.	2. Efficacy	3. Generalizability	4. Cost Benefit	5. Opportunity Cost	6. Fidelity	7. Outcome Assessment	8. Do-ability	9. Implementation	Other:	**Total:**
Scores											

Aggregate Score Sheet for use with the Program Assessment Rubric v1

Scoring: Write your rating in the box under the criteria. When done, add the scores and put the total in the box to the far right. Note the program or proposal being rated below: **Programs/ Proposals or Raters*:**	1. Objective of Program	2. Efficacy	3. Generalizability	4. Cost Benefit	5. Opportunity Cost	6. Fidelity	7. Outcome Assessment	8.Do-ability	9. Implementation	Other:	Total(s):
1.											
2.											
3.											
4.											
5.											
6.											
7.											
8.											
9.											
10.											
11.											
12.											
13.											
14.											
15.											
16.											
17.											
18.											
19.											
20.											

*When using the Program Assessment Rubric v1, this aggregate score sheet can be used to compare different