Prev Sci (2015) 16:933–937
DOI 10.1007/s11121-015-0583-6

The Revised Quality Standards: "A Man's Reach Should Exceed His Grasp" or "A Bridge Too Far": Which is the Case?

Leonard Bickman [1,2]

Published online: 2 August 2015

Introduction

I feel confident, knowing many of the members of the task force who developed the revised standards, that several of the ideas or concerns that I am going to address are not new to them. I am assuming that if my viewpoints are not reflected in the standards' documents, they were either not considered sufficiently important, or they did not achieve the consensus required for a document birthed by a committee. I have tried to capture the theme of my commentary in the title. The dialectic between having aspirational standards that are far reaching versus standards that are so far ahead of the field that may seem unreasonable to obtain.

I will begin this paper by comparing the revised standards to a paper that was published in 1983 entitled "The Evaluation of Prevention Programs" (Bickman 1983). I also ask the reader's forbearance in not looking askance at my reliance on my own references in this paper. Writing this commentary was like a trip down memory lane.

The areas of my concerns in evaluating prevention programs in early 1980 were the following: (a) the almost complete lack of theory to guide program development, implementation, and evaluation; (b) the criteria used to evaluate program success; (c) the evaluation researcher's lack of control over implementation of the program including a need for better theories not only of prevention but also of implementation; (d) the complexity of the chain of causal linkages between program implementation and its ultimate effects; (e) monitoring program integrity over time; (f) measurement of the programs' effects; (g) the establishment of reasonable and meaningful comparison or control groups; and (h) problems of low base rate for establishing sample size. These issues are addressed in the revised standards. These issues, which were not clearly recognized almost 30 years ago, are now an integral part of standards for the next generation. However, I see the implications of some of these standards in a somewhat different perspective than what is expressed in the commentary that accompanied the standards. I describe several of these perspectives in the following text.

Unrealistic Expectations of Degree of Sophistication of Program Theory

As one of the developers of the concept of program theory, I am especially pleased to see that after almost 30 years after publishing a special issue on program theory (Bickman 1987), it is sufficiently recognized that it would be a key ingredient in a set of standards. Program theory is addressed in the following standards: (a) the intervention must be described at a level that would allow others to implement/replicate it, (b) a clear theory of causal mechanisms should be stated, (c) a clear statement of "for whom" and "under what conditions" the intervention is expected to be effective should be stated, and (d) the core components of the intervention and the theory relating these components to the outcomes must be identified and described.

My concern is that the sophistication and advances that have developed in the last 30 years in methodology and statistics far outpaces the progress we have made in program theory. Programs, especially those from non-academics, are probably still more like "notions" than models or theories. While there is an abundance of courses and material on

✉ Leonard Bickman
Leonard.bickman@vanderbilt.edu

[1] Vanderbilt University, Nashville, TN, USA

[2] Florida International University, Miami, FL, USA

methods, statistics, and evaluation, there is a poverty of resources on program development. The most troubling is the first criteria about the details with regard to the intervention. Even if we were able to describe in minute to detail the intervention, we are still not capable of capturing all the key aspects relevant to implementation. We do not know how to identify these key or core aspects of a program. Moreover, the presentation of details may even confuse the attempt at replication because of the impossibility of replicating all the details. These problems are caused by the general absence of theory that identifies the critical factors of the program that must be implemented at some specified level of fidelity to both be considered a replication and to find an effect. One of the most frustrating interchanges I have experienced is trying to get program developers to specify the core components of their program. Components, that if not present with sufficient fidelity, would cause the program to fail. It is not only frustrating for me, but also for the program developers, since they rarely have data or strong theory to support their conclusions.

The standards place little emphasis on the context in which a program is implemented. The description of context is as listed as a desirable standard for effectiveness trials, but what is not emphasized in these or other standards is the inclusion of the program staff in considering standards. Program developers typically include training and sometimes supervision in their program description, but rarely do we find staff descriptions with regard to knowledge, experience, motivation, and abilities. Social and educational programs do not get implemented without humans. But we do not describe staff along those dimensions or place specific requirements for implementation. The lack of attention paid to program personnel has existed for at least 30 years, but no solution appears forthcoming (Peterson and Bickman 1992). For example, we have been struggling over 50 years in trying to determine the relative importance of the therapist vs. the therapy. I think the absence of these specifications in both programs and standards is simply because we do not know the answers. Including staff descriptors, whose contribution to either fidelity of implementation or outcomes is at best unknown, may make already complex situation just more complex. Under these circumstances, it may be best to avoid including them.

The Standard as it Affects Funding

We can define more rigorous and transparent standards as a preventive intervention. If so, what is the implicit program theory of the intervention of increasing rigor? Simply, is it higher standards that lead to better research and evaluations (this is a bit circular)? But what are the mediating links? It is clearly not a direct linkage. Higher standards promulgated by professional association lead to higher standards applied in education, publishing, and in the awarding of grants. Designs and analyses just a few years ago are no longer acceptable in some journals or review panels. In terms of the latter, I have seen the change over the 8 years I served as a peer reviewer on several panels for IES. While points are awarded in several categories (i.e., Significance, Research Plan, Personnel, and Resources), the category that was most often the determining factor of an award was the Research Plan that included methods and statistical analyses. Delving even deeper, it was the statistical power that played a major role in determining funding. There are several possible reasons for this emphasis. First, there are no absolute standards for such categories as significance. However, the statistician on the panel usually spoke with complete authority when he or she said that the statistical power was not sufficient to warrant funding the study regardless of the significance of the study. It is unlikely that IES funded many underpowered designs. Power was just one issue. The designs became more complex and costly as well as demand for better conducted pilot studies and psychometric data. Yes, I agree this is positive but the danger is that we are funding technically superior studies that are not significant. Albert Einstein purportedly warned us, "Not everything that can be counted counts, and not everything that counts can be counted." Significance should count, but it is more difficult to value than calculating statistical power.

The Standard as it Influences Publishing

I am editor-in-chief and founder of a journal that goes by the awkward name of Administration and Policy in Mental Health and Mental Health Services Research. The title is based on the merger of a previously existing journal and the one I founded, Mental Health Services Research that were published by the same publisher. The journal has one of the highest impact factors in the field, so by this metric, it has been successful. However, the review process is burdened by the submission of manuscripts that have significant methodological flaws or lack clarity about the procedures and analyses. As the editor-in-chief, I screen out many of these manuscripts but give the authors the benefit of the doubt when it comes to innovative studies or studies in under-researched areas.

Some journals have required authors to use a standard of evidence, such as the CONSORT checklist for randomized experiments, as a way for authors to screen their own manuscripts before submission. I have had several e-mail exchanges with my associate editors concerning adopting a policy requiring authors to complete a checklist that accompanies the relevant standards. All of us are in favor of such a policy but with reservations. We realize that this will make it even more difficult to submit manuscripts (especially ones that require prior registration of a clinical trial). However, we do

expect that the review process will be more streamlined and that the reviewers' job might be less time consuming when they can see how the author completed the checklist. A review of relevant standards is leading us to adopt at least six different standards. These range from the more familiar CONSORT standards for randomized designs to the less familiar STROBE standards used with observational studies in epidemiology. I feel certain that we also will be adding the Society for Prevention Research Standards of Evidence since they include an interesting and important dimension absent from other standards, namely the purpose of the study as an important dimension (i.e., efficacy, effectiveness, or scale-up). The authors have some flexibility in selecting which standard to use or to provide a rationale as to why none of the published standards are relevant. I think there is no question that standards will affect both what is submitted and what is ultimately published. We will be carefully examining relevant metrics such as number of submissions, publication rate, and review time to see what this intervention has on the journal.

The Attrition Devil (as in the Devil is in the Details)

The standards clearly recognize that attrition, especially differential attrition, is a significant problem in all designs. The standards also highlight that the extent and patterns of missing data must be reported. The committed and diligent researcher can predict with almost certainty that attrition will occur but other than reporting it what should be done? Clearly, the standards are not designed to provide solutions for all the problems that will occur, but in this case, the need for further guidance is more than desirable. Some years back, I collaborated with my good friend, the late Mike Foster, in trying to provide some guidance concerning attrition (Foster and Bickman 1996). However, the Society's standards commentary indicate that current statistical methods (let alone those proposed in 1996) of imputation, which is a key approach to dealing with attrition, is limited, because data rarely meet the assumptions required of such approaches. Is there a solution on the horizon? Can we anticipate that standards can motivate research on how to manage difficult methodological and statistical problems?

The Unintended Consequences of Raising the Bar is to Lower the Boom

The Society's commentary is sensitive to many of the potential negative effects of raising the standards of evidence. The commentary recognizes Voltaire's wise statement "Le mieux est l'ennemi du bien." (The perfect is the enemy of the good). However, adoption of the revised standards may result in some of the following unintended negative outcomes.

Increasing Cost There is no doubt that applying the standards will result in increased costs associated with enhancing the rigor of evidence. I have not agreed with the tactic of a specific budget percentage of a program allocated to an evaluation. My argument is that the amount of funds spent on an evaluation should be commensurate with the evaluation question being addressed and the certainty required for the answer. The cost of the program is only a minor factor in my reasoning. The standards clearly increase the certainty required for an answer. The ramifications are manifold. For example, take the simple case of a training intervention. If we apply the rigor of the standards to pilot testing, then how much pilot work needs to be done to decide on such relatively simple factors as the number of hours of training needed or the spacing of training let alone more complex issues of the content of training. Should the standards be applied to determining what are the core components of programs? If so, the cost of pilot work and planning will rise dramatically. How much evidence should exist to support the mediating linkages in a program theory expressed in a logic model? These issues do not even address the increased costs associated with establishing valid measurement or multiple control groups.

Reducing the Number of Evaluations Unless you live a different world from me, I do not see under the current political conditions in the U.S. that there will be increased funds associated with research and evaluation. Given widespread adoption of the standards, it is simple to conclude that the number of evaluations will be greatly reduced. This may be a good thing for everyone but the evaluators. Maybe fewer evaluations of better quality is a good outcome, but it may mean that evaluations become even rarer, which is not a good outcome.

Devaluing Typical Evaluations Much less than perfect evaluations may still be of value. The choice will not be mediocre evaluations vs. a better evaluation, but mediocre vs. no evaluations. Will the more typical non-federally funded small-scale evaluation be driven out of existence because its flaws are more transparent?

Privileging the Already Privileged Attempting to follow the standards will not only be more expensive, but will also require more skillful and experienced researchers and evaluators. The already advantaged large corporate evaluation companies will be further advantaged.

Increasing Frustration of Stakeholders—Who Cares About Rigor and Transparency? I have defined my professional outlook as a skeptical optimist. I am skeptical about any findings (especially positive ones), but I am optimistic that our scientific methods will uncover the truth. Lately, however, I am becoming more cynical about who really wants more

rigorous methods. Clearly, academics and researchers want them as evident by the plethora of guidelines and checklists. It is a guiding principle of most researchers that the more rigorous the research or evaluation, the more valid the findings. But we should be clear that higher standards do not mean more useful research. As long as researchers only talk to each other, there is relatively harmonious agreement about the need to improve and codify our standards. But what does the rest of the world think?

I do not think there is widespread support for more rigorous standards outside of the research community. I did a quick review to see if there were any data to support my opinion and did not find anything substantial. In the early years of program evaluation (1970s), there was a great deal of strum and drang about evaluation utilization, but that concern seems to have abated. Not that we had a satisfactory answer about utilization, but I think we got bored with questioning the usefulness of our own field. It is much more fun to demand usefulness of other fields. I did find an interesting commentary on the tension between relevance and rigor by Reeves (2011) in the field of education, but no data. Another interesting commentary is also based in the field of education that focused on the What Works Clearinghouse. This author noted "The institute, which serves as the research arm for the Education Department, established the clearinghouse in 2002 in order to showcase high-quality research for educators and policymakers. Yet its rigorous criteria for inclusion, which focuses primarily on randomized, controlled trials and quasi-experimental studies, at first, found relatively few studies that met the bar. Of those that met the criteria, so few showed strong positive effects that the site was given the moniker the 'Nothing Works Clearinghouse' (Sparks 2010)." In terms of sustainability, the Clearinghouse should be considered a success since it has held to its very rigorous standards, with only a little modification, for 13 years.

With the lack of any systematic data to support my opinion, I am left to consider my own idiosyncratic experiences to bolster my argument that it is mainly researchers who are interested in increased rigor. Let us start with a point on which there is widespread agreement. Increased rigor will cost more. That alone may be sufficient to inhibit the growth of rigor. Once we move out of the realm of federal funding and large foundations, the cost/benefit ratio of rigor to cost becomes even more important. Stakeholders have little appreciation for the value of rigor, even if they had a graduate course in research design or statistics. Frontline practitioners maintain even less support for rigorous research. Clinicians do not even like numbers (Bickman et al. 2000). Providers have to live with budgets and demands for services that compete with evaluation dollars.

Although it is far from clear in the literature, it may be that less rigorous designs more often produce statistically significant results. Of course, this would depend on the type of bias introduced into the less rigorous design. For example, the so-called researcher allegiance, where the researcher has ties to the program being evaluated, tends to produce more statistically significant results (Munder et al. 2013). Thus, selected biases can work in favor of positive findings. While RCTs can suffer from researcher bias, they are less likely to have biases associated with such factors as selection. Thus, program developers may benefit from some types of less rigorous research. There is also another assumption that comes into play here. What is the percentage of prevention programs that are truly efficacious, effective, or can be scaled? If less rigorous designs were more likely to produce type II errors, then it would be surprising if non-researchers supported more rigorous designs. However, maybe because I have conducted many evaluations that have not found significant effects, I do not see these outcomes as failures but as opportunities to learn (Bickman and Athay 2009).

Finally, non-researchers developing and/or implementing a program are looking to focus on what counts toward their continued existence, which is usually simple input or output metrics. Complex studies that take several years to conduct simply get in the way of their major job, which is to implement a program and keep their organization funded. If rigorous evaluations can help programs, then fine, but if rigorous standards obstruct programs, then they will only support the less rigorous ones. In the state in which I currently live (Florida), the citizens have the ability to support special taxing districts. These districts include fire and water and children's services. The children's services districts are not government agencies and are only have to be renewed about every decade during a general election to continue their status. Recently, the local newspapers strongly supported the renewal of one of these agencies. The only data the newspaper used to base their support was input data, i.e., how many children they have served. While the agency does do some program evaluations, I would not label this a serious effort, despite their over $100 million a year budget. This personally dismays me since these agencies have the freedom to fund almost any program and thus greatly contribute to filling knowledge gaps. But as long as there is no demand (either internal or external) for more rigorous evaluation efforts, why should programs spend more money and time evaluating their efforts when it does affect their bottom line (i.e., continued existence)? We are most likely to be able to do rigorous evaluations when it required externally or with the small percentage of programs that have enlightened leadership.

Conclusions

It is considerably easier to identify problems, but it is much more difficult to suggest solutions. How do you avoid sounding trite in saying more research is needed? Maybe we can

avoid this fate by being more specific about what research is needed. Moreover, we can try to martial resources by encouraging and supporting the identification and funding of the "more research" through our professional associations, contacts with funding agencies and journals. Despite some limitations, I have noted I support the aspirational approach that extends our grasp. However, unless we convince the non-research community that rigor will be best for them in the long run, we can expect significant push back.

Conflict of Interest The author declares that he has no conflict of interest.

References

Bickman, L. (1983). The evaluation of prevention programs. *Journal of Social Issues, 39*, 181–194.

Bickman, L. (Ed.). (1987). Using program theory in evaluation, new directions for program evaluation, No. 33. San Francisco: Jossey Bass.

Bickman, L., & Athay, M. M. (2009). The worst of all possible program evaluation outcomes. In A. R. Stiffman (Ed.), *The field research survival guide* (pp. 174–204). New York: Oxford University.

Bickman, L., Rosof, J., Salzer, M. S., Summerfelt, W. T., Noser, K., Wilson, S. J., & Karver, M. S. (2000). What information do clinicians value for monitoring adolescent client progress and outcome? *Professional Psychology: Research and Practice, 31*, 70–74.

Foster, E. M., & Bickman, L. (1996). An evaluator's guide to detecting attrition problems. *Evaluation Review, 20*, 695–723.

Munder, T., Brutsch, O., Leonhard, R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: An overview of reviews. *Clinical Psychology Review, 33*, 501–511.

Peterson, K. A., & Bickman, L. (1992). Using program theory in quality assessments of children's health services. In H. Chen, P. Rossi (Eds.), *Using theory to improve program and policy evaluations* (pp. 165–176). Westwood, MA: Greenwood Press.

Reeves, T. (2011). Can educational research be both rigorous and relevant? *Educational Designer, 1*(4), 1–24.

Sparks, S. D. 2010." 'What Works' Broadens Its Research Standards: Clearinghouse Moves Past 'Gold Standard.' " *Education Week, 30*(8), 1, 12.